DOI: 10.1002/ev.20611

ORIGINAL ARTICLE

WILEY

Relational Systems Evaluation rubrics as tools for building and measuring evaluation capacity

Jennifer Brown Urban[1] | Elyse L. Postlewaite[1] | Bekki Davis[1] | Elaine Les[1] | Jane Buckley[2] | Monica Hargraves[3]

[1] Institute for Research on Youth Thriving and Evaluation, Montclair State University, Montclair, New Jersey, USA

[2] Institute for Research on Youth Thriving and Evaluation, Independent Consultants, West Henrietta, New York, USA

[3] Institute for Research on Youth Thriving and Evaluation, Independent Consultants, Ithaca, New York, USA

Correspondence
Jennifer Brown Urban, Montclair State University, Montclair, NJ, USA.
Email: urbanj@montclair.edu

Abstract

Evaluation capacity building (ECB) is still an emerging area of study in the field of evaluation. The purpose of ECB is to assist program practitioners with implementing higher-quality evaluation; however, we need better tools and resources to effectively assess ECB efforts. Existing measures typically depend on self-report as opposed to assessing the artifacts of ECB training. Among the few non-self-report tools that support the assessment of ECB efforts are Relational Systems Evaluation rubrics designed to evaluate logic models, pathway models, and evaluation plans. These rubrics were first developed and tested several years ago. The purpose of the current study is to update the Relational Systems Evaluation rubrics to reflect current ECB knowledge. The updated rubrics have good to excellent inter-rater reliability and high internal consistency. The results of this study contribute to the ECB field by providing measurement tools for assessing the quality of ECB artifacts. The rubrics can also be used by organizations and funders who need a systematic approach for assessing (and comparing) the quality of evaluation plans and visual theory of change models (e.g., logic models).

Jennifer Brown Urban and Elyse L. Postlewaite contributed equally to this study.

© 2024 American Evaluation Association and Wiley Periodicals LLC.

INTRODUCTION

Conversations around what constitutes high-quality evaluation are not new and have evolved over time. In addition to having technical quality, high-quality evaluation must also be beneficial, sustainable, democratic, culturally responsive, and well-aligned (e.g., with resources, context, theory of change, program lifecycle; Hood et al., 2014; House, 2019; Patton, 1982, 2008; Urban et al., 2024). Evaluation quality has been discussed in the context of evaluation capacity building (ECB), yet there are limited tools available for assessing the success of ECB efforts (Labin et al., 2012; Wingate et al., 2022). Most ECB efforts aim to address cognitive (e.g., knowledge, understanding), behavioral (e.g., developing visual theories of change), and affective (e.g., positive attitude) objectives related to evaluation (Preskill, 2008). Self-report measures exist for assessing these objectives (e.g., Bourgeois, et al., 2013; Taylor-Ritzer et al., 2013). Theoretically, successful ECB efforts should also be associated with high-quality evaluation products. Although evaluation products could include the results of a successfully implemented evaluation, a more intermediary assessment of ECB efforts could examine the quality of evaluation planning products, such as visual theories of change (e.g., logic models) and evaluation plans. Urban and colleagues (2015) developed a set of tools for assessing the quality of evaluation artifacts, including evaluation plans, logic models, and pathway models (a particular type of visual theory of change that emphasizes causal connections).

The purpose of this article is to update these quality assessment tools and assess their ability to detect differences between products often developed as part of ECB efforts versus without ECB efforts. We also discuss potential applications of these tools beyond their use in assessing the quality of ECB efforts. This article (1) provides a synthesis of the literature that informed the updates to the rubrics, (2) describes the process of updating and testing the rubrics, (3) assesses whether the rubrics can be used to identify differences in the quality of evaluation products developed by those who engaged in ECB versus those who did not, and (4) discusses how the rubrics can be used to support ECB and other assessment.

ECB

ECB encompasses building knowledge, skills, and attitudes to help program practitioners engage in effective evaluation (Sarti et al., 2017). Organizations' motivation to enhance ECB can stem from internal factors within the organization, such as leadership's ambition to enhance evaluation practices, external factors such as funding requirements, or a blend of both internal and external influences (Labin et al., 2012). Therefore, ECB initiatives and assessments are needed to ensure practitioners have the capacity to implement high-quality evaluations. Workshops facilitated by groups with evaluation experience have been a popular strategy in equipping program staff with knowledge and tools to carry out evaluations (Adams & Dickinson, 2010; Archibald et al., 2016, 2018). Workshops can be effective in providing both higher-level background knowledge and opportunities for facilitated practice that can be subsequently applied to individual and unique programs. Other strategies in training and supporting individuals with nonevaluation backgrounds have come to the forefront. One strategy is providing consistent feedback to evaluators in the form of one-on-one and team meetings, open lines of communication, and a "buddy"-type system between more and less experienced evaluators (Rogers et al., 2018). Similarly, partnerships between members of the organization and evaluation practitioners rooted in open and consistent communication can add beneficial support at all stages of the evaluation (O'Brien et al., 2017). Furthermore, it can be beneficial for program staff to partner with an evaluation facilitator (e.g., an internal evaluator or external consultant) so they can

1534875x, 2024, 183, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ev.20611 by <Shibboleth>-member@80182490.imf.org, Wiley Online Library on [25/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

ask questions, receive feedback, and have access to evidence-based materials to support effective evaluation (Buckley et al., 2024).

ECB assessment tools

ECB is a relatively new concept in evaluation (Bourgeois et al., 2023). As such, there are few empirically validated tools that assess the outcomes of ECB efforts (Taylor-Ritzer et al., 2013). The tools that exist often rely on self-assessment measures (Bourgeois et al., 2023). Two of the better-known self-report measures of ECB are the evaluation capacity assessment instrument (Taylor-Ritzer et al., 2013) and the organizational evaluation capacity self-assessment instrument (Bourgeois et al., 2013). The evaluation capacity assessment instrument is an empirically validated measure for examining evaluation capacity in nonprofit organizations. The organizational evaluation capacity self-assessment instrument measures an organization's ability to both "do" and "use" evaluation in six key areas and is completed by a collaborative team of evaluation staff and managers.

In addition to self-report measures, objective or observational measures of ECB can be useful when self-report measures cannot be used or will not elicit appropriate information. Other assessment tools, such as rubrics, can elicit more objective information about the quality of an evaluation (Urban et al., 2015). Self-report instruments are prone to bias and error (Donaldson & Grant-Vallone, 2002). For example, some aspects of evaluation capacity, such as the quality of evaluation logic models, may be difficult for individuals to self-assess, especially if they are novices. Recognizing this difficulty and after identifying a need for evaluation research that used tools other than self-reports, Wingate and colleagues (2022), created an observational rubric allowing independent investigators to assess the inclusion of six key evaluation factors in evaluation plans; this rubric was adapted from the evaluation plan rubric developed by Urban and colleagues (2015).

The Urban et al. (2015) rubrics were developed to both provide feedback to participants engaged in ECB initiatives and assess the efficacy of capacity-building efforts using Relational Systems Evaluation and the Systems Evaluation Protocol (Urban, Archibald, et al., 2021, Urban, Hargraves, et al., 2021; Urban et al., 2015). Since the creation of the original Urban et al. (2015) rubrics, new theories and approaches to evaluation have gained prominence, and therefore the rubrics would benefit from being updated to include current evaluation concepts.

Relational systems evaluation

Relational Systems Evaluation is a theoretically grounded and empirically tested framework for program planning and evaluation that centers the relationship between evaluation facilitator and program practitioner; draws on concepts from systems science, relational developmental systems meta-theory, and evolutionary epistemology to guide the evaluation approach; and emphasizes evaluative thinking in building evaluation capacity (Urban, Hargraves, et al., 2021). Relational Systems Evaluation is theoretically grounded in evolutionary theory (Darwin, 1859; Mayr, 2001), evolutionary epistemology (Bradie & Harms, 2006; Campbell, 1974, 1988; Cziko & Campbell, 1990; Popper, 1973, 1984), systems theory (Bertalanffy, 1995; Laszlo, 1996; Midgley, 2003; Ragsdell et al., 2002), and relational developmental systems meta-theory (Lerner, 2006; Overton, 2006, 2010). Relational Systems Evaluation was developed, refined, and tested over 15 years through a series of projects in a variety of disciplines/fields (e.g., Cooperative Extension, National Science Foundation—funded education outreach programs, youth development programs).

1534875x, 2024, 183, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ev.20611 by <Shibboleth>-member@80182490.imf.org, Wiley Online Library on [25/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

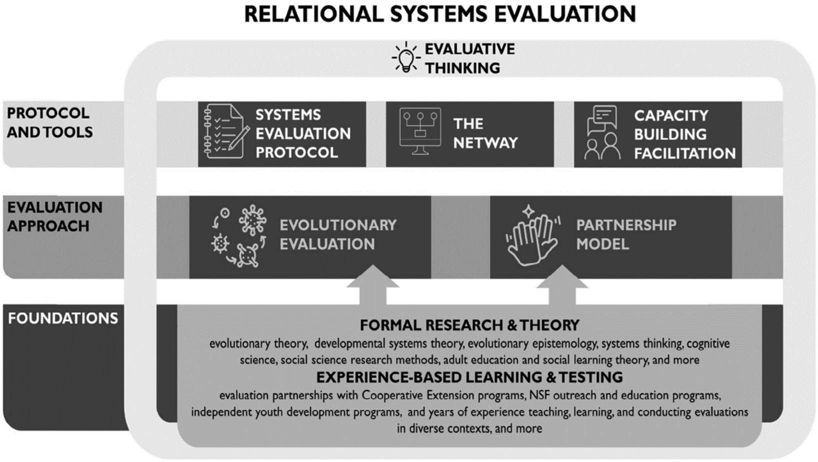
RELATIONAL SYSTEMS EVALUATION
EVALUATIVE THINKING
PROTOCOL AND TOOLS
SYSTEMS EVALUATION PROTOCOL
THE NETWAY
CAPACITY BUILDING FACILITATION
EVALUATION APPROACH
EVOLUTIONARY EVALUATION
PARTNERSHIP MODEL
FOUNDATIONS
FORMAL RESEARCH & THEORY
evolutionary theory, developmental systems theory, evolutionary epistemology, systems thinking, cognitive science, social science research methods, adult education and social learning theory, and more
EXPERIENCE-BASED LEARNING & TESTING
evaluation partnerships with Cooperative Extension programs, NSF outreach and education programs, independent youth development programs, and years of experience teaching, learning, and conducting evaluations in diverse contexts, and more

FIGURE 1 Components of Relational Systems Evaluation.

A key component of Relational Systems Evaluation is an evolutionary evaluation approach that defines rigorous evaluation as including the goodness of fit between a program's lifecycle phase and the evaluation's methodological design. Alignment between the program and evaluation lifecycles is essential for ensuring that the evaluation will be efficient, fiscally responsible, and useful. A lack of alignment (e.g., using a randomized controlled trial to evaluate a program that is being implemented in the real world for the first time) risks a poor use of resources and misleading evaluation results (Urban et al., 2014).

Relational Systems Evaluation is implemented in a facilitated partnership model, including partners with expertise in evaluation and those with expertise in the focal program, context, and participant population. The relational partnership model focuses on responsiveness, trust building, and mutual respect, with the accompanying goal of elevating practitioner voice (Buckley et al., 2021).

Due to the focus on evaluation in the context of facilitated research/evaluation-practice partnerships, Relational Systems Evaluation is inherently an ECB method and can be facilitated through the implementation of the Systems Evaluation Protocol. The Systems Evaluation Protocol is both a flexible protocol and a conceptual framework designed to facilitate thinking about the complex systems that influence evaluation planning decisions. It includes a step-by-step guide for developing an evaluation plan and visual theory of change (Trochim et al., 2016; Urban, Hargraves, et al., 2021). The Netway is a free online platform that includes an electronic version of the Systems Evaluation Protocol, instructional videos, worksheets, and tools that aid in developing evaluation plans and visual theories of change. In Relational Systems Evaluation, we encourage partners to develop a particular type of visual theory of change called a pathway model. Pathway models are visual diagrams that include components of columnar logic models (i.e., activities and outcomes) and also specify the hypothesized causal connections between specific activities and outcomes, thereby communicating a detailed version of the theory of change underlying the program. Figure 1 provides a visual overview of the components of Relational Systems Evaluation.

How the original rubrics were developed and tested

Relational Systems Evaluation and the Systems Evaluation Protocol have been implemented using three distinct approaches. The "bootcamp" approach involves an intensive

1534875x, 2024, 183, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ev.20611 by <Shibboleth>-member@80182490.imf.org, Wiley Online Library on [25/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

partnership between a Systems Evaluation Protocol facilitator (evaluator) and a small group of program stakeholders (typically including practitioners, organizational leaders, and/or program developers). The "practitioner training" approach focuses on using Relational Systems Evaluation and the Systems Evaluation Protocol for ECB and typically involves a series of workshops led by Systems Evaluation Protocol facilitators. Workshop attendees include two to three representatives from each of 5 to 10 participating programs. Often the participating organizations are all part of a larger system or network of programs (e.g., grantees who are all funded as part of the same portfolio or initiative). The workshops are followed by tailored consultations between a Systems Evaluation Protocol facilitator and members of each program. A third approach is "independent implementation," whereby evaluators or program practitioners access the Systems Evaluation Protocol materials and resources through the Netway without facilitation (Urban et al., 2018). The Netway contains artifacts, including evaluation plans, logic models, and pathway models, from 3771 programs. Many of these artifacts are from programs that engaged in facilitated evaluation partnerships (i.e., bootcamp or practitioner training); however, a substantial portion of artifacts are from programs that did not have any interaction with System Evaluation Protocol facilitators.

As ECB facilitators implementing the "practitioner training" approach, we developed rubrics for evaluation plans, logic models, and pathway models that allowed us to provide systematic feedback to our partners. We later adapted the rubrics so they could be used to score the quality of evaluation planning products. In 2015, we published the rubrics as a tool for assessing the effectiveness of ECB (Urban et al., 2015). Since the original publication of the rubrics, the field of evaluation has evolved, and we have gained more experience implementing Relational Systems Evaluation. As such, we felt it was time to update the rubrics to reflect these knowledge gains. In the next section, we describe recent advances that informed our revisions.

UPDATES TO THE RUBRICS

Based on our review of the literature, we identified several best practices for assessing the quality of evaluation plans and visual theory of change models. Our literature search specifically focused on contemporary concepts in evaluation aligned with the values of the research team (e.g., culturally responsive evaluation) and sought checklists and existing rubrics. In addition, we reviewed and synthesized resources, tools, and protocols developed by and representing the internal knowledge and expertise of the research team.

Contemporary evaluation concepts reviewed

We identified several core concepts and paradigms that were missing from the original rubrics, including transformative evaluation (Mertens, 2009); culturally responsive evaluation, including culturally appropriate and accessible data collection, customizing reporting methods to fulfill communication and language needs, reflexivity, and critical self-reflection of the evaluation team (Hood et al., 2014; Hopson, 2003); AI and using innovative technology with a critical lens (Tilton et al., 2023); and character-minded evaluation (Urban et al., 2024).

Transformative evaluation

Transformative evaluation is a paradigm that situates social justice and human rights at the core of evaluation. Key themes of this paradigm include underlying assumptions that

1534875x, 2024, 183, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ev.20611 by <Shibboleth>-member@80182490.imf.org, Wiley Online Library on [25/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

challenge oppressive structures, an entry process designed to build trust with the community, and dissemination of findings to encourage use for social justice efforts (Mertens, 2009). Transformative evaluation draws from feminist theories, critical race theory, queer theory, intersectionality, and postcolonial and Indigenous theories and emphasizes the importance of forming authentic relationships, being culturally responsive, and building coalitions to enhance ownership and utilization of the evaluation. The transformative evaluation paradigm can serve as the foundation for an evaluation model, theory, or approach. Within this paradigm, evaluators use diverse methodological approaches ranging from appreciative inquiry to community-based and participatory methods. These methodological approaches facilitate collaboration among populations, organizations, and evaluators while emphasizing trust, transparency, and attention to power differences (Bolinson & Mertens, 2019). Ultimately, transformative evaluation aims to support the type of change needed to address complex societal challenges and work toward a more equitable and just world.

Culturally responsive evaluation

Culturally responsive evaluation is rooted in centering the culture of the program being evaluated (Frierson et al., 2010; Hood et al., 2015). It calls for incorporating culture into evaluation to serve the specific, unique, and diverse characteristics of the evaluation context. Culture should be integrated through all stages of evaluation in theory and practice (Hood et al., 2014). A key component of practicing culturally responsive evaluation is ensuring that evaluators are knowledgeable about the population and culture of participants in the program and those impacted by the evaluation. This includes facilitating the selection and implementation of appropriate measures (Hood et al., 2014). Combining core concepts of culturally responsive assessment and pedagogy and responsive evaluation, culturally responsive evaluation pays special attention to groups who have been historically marginalized (Hopson, 2009). Culturally responsive evaluation is not a set of tangible steps for evaluators to follow; instead, cultural competence must be embedded in each step of the evaluation lifecycle. Culturally responsive evaluation has been embedded throughout the updated version of our evaluation plan rubric, specifically, for example, by calling for attention to the cultural context of the program and target audience and ensuring that data is reported in ways that are accessible and culturally responsive.

Artificial intelligence

Artificial intelligence plays a pivotal role in evaluation education and ECB (Head et al., 2023; Tilton et al., 2023). Generative AI can increase access to meaningful learning opportunities and experiences that were traditionally only accessible through postsecondary education (Arévalo Gross et al., 2022). Generative AI is also an interactive tool in which evaluators can share or “discuss” ideas or concepts. Tilton and colleagues (2023) outline an example in which students might formulate a program theory, then interact with an AI chatbot to identify social science theories that pertain to the specified short-, medium-, and/or long-term outcomes outlined in the program’s logic model. Subsequently, students can engage the chatbot agent to draw inferences and highlight the assumptions inherent in the logic of program theories. AI chatbots can also be helpful in pinpointing or identifying unexpected issues with artifacts of ECB, such as logic models or evaluation plans. While the benefits of integrating AI into ECB are exciting, there are noteworthy risks and potential drawbacks. Foremost, prompts used with AI platforms need to be well crafted and highly

1534875x, 2024, 183, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ev.20611 by <Shibboleth>-member@80182490.imf.org, Wiley Online Library on [25/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

specific, and users may need training and practice in order to generate the type of output they seek (Head et al., 2023). Large language models can also be factually inaccurate and perpetuate biases, reinforcing the importance of using AI and large language model tools with a critical lens. Updates to the evaluation plan rubric include references to AI, specifically encouraging evaluators and practitioners to use a critical lens and transparency in reporting AI use to support evaluation efforts.

Character-minded evaluation

Character-minded evaluation is defined as "evaluation work that intentionally promotes and draws on the power of character virtues and practical wisdom to develop high-quality evaluation and flourishing programs" (Urban et al., 2024, p. 310). Character-minded evaluation accounts for the greater good, multiple perspectives, and the ultimate usefulness of evaluation results, while evaluation work that is not character-minded tends to result in unexplored data, disregarded reports, and stagnant program strategies. A central idea in character-minded evaluation is that there is a "bi-directional relationship between character and high-quality evaluation: character virtues and practical wisdom contribute to good evaluation *and* good evaluation practices contribute to character virtues and practical wisdom" (Urban et al., 2024, p. 311).

Character-minded evaluation applies the Jubilee Centre's Framework for Character Education in Schools (Jubilee Centre for Character and Virtues, 2022). According to this framework, individual character virtues can only be operationalized effectively when integrated with practical wisdom. Practical wisdom is "the integrative virtue" and allows us to discern which virtues to apply and how to act when "virtues collide" (Jubilee Centre for Character and Virtues, 2022). Practical wisdom is what allows evaluation planners and implementers to stay aligned with the civic virtues that serve as the ultimate purpose for their evaluation work, even as they make the many small decisions that go into planning an evaluation (Hurteau & Archibald, in press; Schwandt, 2015; Schwandt & Gates, 2021; Urban et al., 2024). In the rubrics presented here, there is an emphasis on alignment to the stated purpose of the evaluation and its guiding questions. The items related to alignment allow users of the rubrics to acknowledge evidence of the application of practical wisdom as well as alignment with the broader civic, moral, intellectual, and performance virtues that underlie the evaluation's purpose.

Reflective conversations to update rubrics

After completing the literature review, we made updates to the rubrics through a series of interactive reflective conversations among the six research team members. For each rubric, we took turns with one member of the research team providing suggested updates and the whole team responding to the suggestions. In the reflective conversations, we also analyzed the rubrics for comprehensiveness and comprehensibility. Individual team members raised issues, and we discussed them until we reached consensus. For further refinement, all team members completed initial scoring of program artifacts (see sample section)—including 30 visual theory of change models and nine evaluation plans—using the updated rubrics.

Evaluation plan rubric

As part of the refinement process, we made a slight change to the scoring scale: 0 now means *absent/missing*, whereas in the original rubric 0 meant *unacceptable*. We added

1534875x, 2024, 183, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ev.20611 by <Shibboleth>-member@80182490.imf.org, Wiley Online Library on [25/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

items and updated existing items to include the evaluation concepts reviewed above. We also clarified language within the items, and then removed four items for redundancy. Finally, we collapsed two subsections of the rubric into one to better represent the focal area of the items.

Visual theory of change rubric

Reflective conversations led to combining the original pathway model rubric and logic model rubric into one visual theory of change rubric. The revised rubric is more flexible and can accommodate the many formats used to depict programmatic theories of change, including more traditional columnar logic models as well as more detailed pathway models.

We adjusted the structure of the visual theory of change rubric to provide greater clarity to users and to ensure the sections were mutually exclusive. The final rubric structure includes four sections that address (1) the quality of the model structure, (2) the quality of the nodes (boxes), (3) the quality of the links (arrows/connections), and (4) the quality of the overall model as a theory of change tool. We kept the scoring scale for the visual theory of change rubric the same as the original logic model and pathway model rubrics.

CURRENT STUDY

After making initial updates to the rubrics, we turned our focus to refining and testing the revised rubrics. Unobtrusive measures of evaluation planning artifacts should provide an early indicator of whether ECB efforts are successful. If evaluation capacity has been successfully built, evaluation artifacts should be of high quality. Our empirical work aims to assess whether rubrics can be used as a proxy for measuring ECB effectiveness, addressing two research questions: (1) Are the revised Relational Systems Evaluation rubrics reliable? And (2) Do programs that engaged in facilitated ECB efforts have higher-quality visual theories of change than programs that have not engaged in facilitated ECB?

METHODS AND ANALYSIS

Research design

This study used a mixed methods design to score evaluation artifacts (visual theory of change models and evaluation plans) and test rubric reliability. A nonequivalent group post-only quasi-experimental design was used to test the hypothesis that programs that engaged in facilitated ECB efforts would have higher-quality visual theories of change than programs that have not engaged in facilitated ECB.

Sample

Visual theory of change models

An initial sample of visual theory of change models was identified in the Netway platform ($n = 3771$). The visual theory of change models were reviewed for completeness, including having at least one entry in each category: activities, short-term outcomes, mid-term

1534875x, 2024, 183, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ev.20611 by <Shibboleth>-member@80182490.imf.org, Wiley Online Library on [25/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

outcomes, and long-term outcomes. Programs were excluded from the sample if they were not written in English. This brought the working sample to $n = 2102$.

Three of the six research team members (Urban, Hargraves, and Buckely) had previously provided ECB facilitation using either the Relational Systems Evaluation "bootcamp" or "practitioner training" approach with some of the programs in the sample. Programs that had been through one of these facilitated processes were identified and labeled as "facilitated." Programs that had not received guidance or expertise from the research team in the development of the visual theory of change model were identified and labeled as "non-facilitated." Programs whose status (having gone through a facilitated process or not) we were unsure about were excluded from the sample. A total of 30 models were randomly selected using a random number generator that were used for the reflective conversations and initial updates to the rubric. A total of 60 models were subsequently randomly selected using a random number generator (30 facilitated models and 30 nonfacilitated models) that were used for the reliability analyses.

Evaluation plans

We gathered an initial sample of 68 evaluation plans from the following sources: plans created by students as part of evaluation-related academic coursework, evaluation plans available in the Netway platform, and evaluation plans published on the Internet. To locate Internet-published plans, team member Davis used the following search terms with a Chrome web browser: "evaluation plans," "evaluation plans in grants," "program evaluation plans public health," "evaluation plans in education," "evaluation plans in nonprofit." Plans were then downloaded from state and federal websites, organizations such as departments of education and schools, and nonprofit organizations. After one round of applying the evaluation rubric and discussion, evaluation plans focused on policy development were excluded. This brought the working sample to 57. A total of nine evaluation plans were purposefully selected for scoring. The plans developed by students as part of evaluation-related academic coursework were selected first. Additional plans were added to the study one at a time until sufficient inter-rater agreement was reached.

Rubric reliability analyses

Once initial revisions to the rubrics were completed, all team members' scores were compared for interrater agreement using percent agreement calculations, Pearson's R correlations, Cohen's kappa, and intraclass correlation coefficients (ICC). This process led to several additional changes to the rubrics, primarily to adjust wording to resolve differences in raters' interpretations of the items. The team discussed these changes until consensus was reached.

Once all edits were completed, interrater reliability and internal consistency were assessed to test for consistency among raters and consistency within the rubrics, respectively. Each of the six team members scored six visual theory of change models and nine evaluation plans using the corresponding rubric. The ICC were calculated to determine interrater reliability between randomly paired raters. Cronbach's alpha, inter-item correlations, and corrected item—total correlations were also calculated to verify internal consistency within the rubrics.

1534875x, 2024, 183, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ev.20611 by <Shibboleth>-member@80182490.imf.org, Wiley Online Library on [25/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Rubric hypothesis testing

To address the second research question, we assessed the quality of the visual theory of change models using rubric scores. We hypothesized that programs that had participated in facilitated ECB would have higher scores on the visual theory of change model rubric (a proxy for higher levels of evaluation capacity) than programs that had not engaged in facilitated ECB.

While impossible to completely eliminate potential bias due to familiarity with the programs, the “facilitated” and “non-facilitated” labels were temporarily removed from the dataset to reduce potential bias. After the labels were removed, 60 program models were each randomly assigned to one of five raters and scored. After the visual theory of change models were scored, an independent samples *t*-test was conducted to assess differences in rubric scores between facilitated and nonfacilitated theory of change models.

RESULTS

Rubric reliability

The results of the reliability analyses are reported separately for the evaluation plan and visual theory of change rubrics.

Evaluation plan rubric

The final evaluation plan rubric includes 49 items divided into 11 subsections: (1) Program Description (five items), (2) Evaluation Purpose (five items), (3) Evaluation Questions (five items), (4) Design (five items), (5) Measures (six items), (6) Sampling (six items), (7) Data Collection and Management (three items), (8) Data Analysis (two items), (9) Evaluation Reporting and Utilization (five items), (10) Evaluation Timeline (three items), and (11) Overall Quality (four items). Rater scores on the evaluation plan subsections were summed to calculate the ICCs. The average interrater reliability score between random pairs of coders was ICC (2, 1) = 0.922, and scores ranged from ICC (2, 1) = 0.854 ($p < 0.01$) to ICC (2, 1) = 1.000 ($p < 0.001$). ICC scores between 0.75 and 0.9 are considered to have good reliability, and scores above 0.9 are considered excellent reliability (Koo & Li, 2016). All ICC values were significant at $p < 0.01$.

Internal consistency for the evaluation plans was calculated using Cronbach’s alpha, inter-item correlations, and corrected item-total correlations. The overall internal consistency estimate for the evaluation plan rubric was calculated using random single-rater scores ($\alpha = 0.950$). See Appendix A for the evaluation plan rubric.

Visual theory of change rubric

The final visual theory of change rubric includes 12 items divided into 4 subsections: (1) Structure (four items), (2) Nodes (four items), (3) Links (three items), and (4) Overall Quality (one item). The ICC scores for the visual theory of change model rubric were also based on the summed subsections of the rubric. The average interrater reliability score between random pairs of coders was ICC (2, 1) = 0.934, and scores ranged from ICC (2, 1) = 0.756 (p = 0.196) to ICC (2, 1) = 0.990 ($p < 0.01$).

1534875x, 2024, 183, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ev.20611 by <Shibboleth>-member@80182490.imf.org, Wiley Online Library on [25/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

1534875x, 2024, 183, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ev.20611 by <Shibboleth>-member@80182490.imf.org, Wiley Online Library on [25/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Internal consistency was also calculated for the visual theory of change rubric. The overall internal consistency estimate for the visual theory of change rubric was calculated using the independently scored models ($n = 60$; $\alpha = 0.871$). See Appendix B for the visual theory of change rubric.

Predictive validity for the visual theory of change rubric

To test the effectiveness of the visual theory of change rubric as an ECB measurement tool, rater scores for the visual theory of change models were compared between facilitated and nonfacilitated models. Results demonstrated a significant difference between facilitated and nonfacilitated models ($t(58) = 4.79$, $p < 0.001$), with facilitated models ($M = 25.67$, $SD = 5.74$) resulting in higher rater scores on average compared to nonfacilitated models ($M = 18.13$, $SD = 6.42$).

DISCUSSION

Given the increased focus on ECB and the rapid development of knowledge related to ECB, rubrics that can be used to assess the quality of evaluation artifacts are needed. The current article described rubrics that can assess the quality of two ECB artifacts: evaluation plans and visual theory of change models. These rubrics were updated from their original versions (developed in 2015 by Urban and colleagues) to include key best practices from contemporary literature as well as the authors' accumulated expertise in the field. Key concepts and paradigms that were missing from the original rubrics and were considered in these updates included transformative evaluation (Mertens, 2009); culturally responsive evaluation, including culturally appropriate and accessible data collection, customizing reporting methods to fulfill communication and language needs, reflexivity, and critical self-reflection of the evaluation team (Hood et al., 2014; Hopson, 2003); AI and using innovative technology with a critical lens (Tilton et al., 2023); and character-minded evaluation (Urban et al., 2024).

Once revised, the rubrics were tested for inter-rater reliability and internal consistency. Both rubrics, the evaluation plan rubric and the visual theory of change rubric, had good to excellent inter-rater reliability and high overall internal consistency. In addition, the visual theory of change rubric was effective at distinguishing between rubrics that were developed as part of a facilitated compared to a nonfacilitated process. Theory of change models created with ECB support had significantly higher scores than those created without ECB support. Within any given rubric, we did not expect the subsection scores to be consistent with one another, because they are theoretically independent. However, we expected internal inconsistency across all items within a particular rubric, and we found this for both rubrics. Given that the rubrics are designed to be modified for specific contexts, we suggest future users go through a process to achieve reasonable inter-rater reliability and internal consistency.

Findings from the present study have the potential to contribute to the ECB field in several ways. First, the updated rubrics can be integrated into already popular ECB training strategies such as workshops in which evaluation trainers connect members of organizations with key tools to support evaluation. The updated rubrics can also help to facilitate meaningful and productive conversation within partnerships between evaluators and members of an organization by completing and discussing the rubric results collaboratively (O'Brien et al., 2017). Bourgeois and colleagues (2023) note the majority of ECB tools used by practitioners are written materials, such as toolkits or manuals. The Urban et al. (2015) rubrics provide a relevant tool that can also be used by practitioners. Furthermore, the current study has improved upon the existing Urban et al. (2015) rubrics by incorporating the evaluation field's important advances that respond to the complexity of the environments surrounding organizations and programs, the commitment to equity and cultural inclusiveness, and the potential contribution of technological innovations. These make the updated rubrics more relevant, both as guides for offering feedback to participants in ECB efforts, and as measurement tools for assessing ECB progress, whether facilitated by ECB practitioners or by an organization's internal evaluation team, or as a self-guided capacity-building effort.

The direct application of these rubrics as tools for assessing the theory of change models or evaluation plans makes them useful to a variety of potential stakeholders. Organization leaders could use the rubrics to systematically assess the quality and/or potential value of evaluation for particular programs and allocate evaluation or programmatic resources accordingly–an identified area of need in ECB intervention literature (Bourgeois et al., 2023). Grantmakers reviewing a portfolio of proposals could use the rubrics to assess the quality of evaluations that are being proposed, the evaluation capacity represented by the applicants, or even the quality of past evaluative work conducted by applicants. By enabling consistent assessment across a set of programs, the rubrics can contribute to both internal and external decision-making and targeted program support. One challenge we encountered at points in our internal scoring rounds was that individuals with different evaluation backgrounds sometimes have different expectations or standards for some components of an evaluation plan or theory of change model. We found that reflective group conversations helped alleviate this (and informed some of our wording changes) and helped us come to a consensus. Moreover, regularly meeting as a group to problem-solve and tap into team members' strengths has proven to be a beneficial strategy for building tools and resources to support ECB (Danseco, 2013).

The rubrics were designed to be comprehensive, identifying a full set of expectations for high-quality evaluation artifacts. Indeed, as noted in the opening instructions in the evaluation plan rubric, users in some contexts may find that the rubric covers some topics or details that extend beyond what is expected in their particular context. Users in such situations are encouraged to customize the tool for their particular situation. This design element highlights the instructive potential of the evaluation plan rubric; because it covers so much, and in considerable detail, it offers a blueprint for achieving high-quality evaluation. This applies as well to the theory of change rubric since having a sound theory of change is foundational for any evaluation work. This instructive use of both rubrics can benefit anyone engaged in evaluation and may be of particular use to organizations and practitioners with limited ECB resources. Thoughtful use of these rubrics can itself build evaluation capacity and contribute toward closing the gap between science and practice by equipping evaluation practitioners with evidence-based measurement tools (Suarez-Balcazar & Taylor-Ritzler, 2014).

While the rubrics can be used in a variety of different applications, it is important to note that they cannot currently be used as predictive measures of other outcomes, such as evaluator skills or evaluation quality. Future research could validate the rubrics so they could be used predictively. Limitations notwithstanding, the current article demonstrates that the rubrics are comprehensive and effective tools that expand the field of ECB in assessment and capacity building.

ORCID

Jennifer Brown Urban https://orcid.org/0000-0001-8405-3078
Elyse L. Postlewaite https://orcid.org/0000-0002-0453-4371
Bekki Davis https://orcid.org/0009-0004-3843-1575

1534875x, 2024, 183, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ev.20611 by <Shibboleth>-member@80182490.imf.org, Wiley Online Library on [25/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Elaine Les https://orcid.org/0009-0001-2404-5233
Jane Buckley https://orcid.org/0000-0003-4206-2341
Monica Hargraves https://orcid.org/0000-0002-6074-4445

REFERENCES

Adams, J., & Dickinson, P. (2010). Evaluation training to build capability in the community and public health workforce. *American Journal of Evaluation*, *31*(3), 421–433. https://doi.org/10.1177/1098214010366586

Archibald, T., Sharrock, G., Buckley, J., & Cook, N. (2016). Assumptions, conjectures, and other miracles: The application of evaluative thinking to theory of change models in community development. *Evaluation and Program Planning*, *59*, 119–127. https://doi.org/10.1016/j.evalprogplan.2016.05.015

Archibald, T., Sharrock, G., Buckley, J., & Young, S. (2018). Every practitioner a "knowledge worker": Promoting evaluative thinking to enhance learning and adaptive management in international development. *New Directions for Evaluation*, *158*, 73–91. https://doi.org/10.1002/ev.20323

Arévalo Gross, C. J., Rodríguez-Bilella, P., & Olavarría, C. (2022). How to train better in evaluation: Teaching landscape and lessons learned from Latin America. *American Journal of Evaluation*, *44*(2), 282–292. https://doi.org/10.1177/10982140211059373

Bertalanffy, L. V. (1995). *General system theory: Foundations, development, applications*. George Braziller Publishers.

Bolinson, C., & Mertens, D. (2019). *A transformative evaluation toolkit for the impact investing sector: How entrepreneurs and impact investors can deepen their data collection*. Engineers Without Borders Canada. Retrieved from https://www.ewb.ca/wp-content/uploads/2019/10/Tool-Kit-document-Interactive_FINAL_Oct31.pdf

Bourgeois, I., Lemire, S. T., Fierro, L. A., Castleman, A. M., & Cho, M. (2023). Laying a solid foundation for the next generation of evaluation capacity building: Findings from an integrative review. *American Journal of Evaluation*, *44*(1), 29–49. https://doi.org/10.1177/10982140221106991

Bourgeois, I., Toews, E., Whynot, J., & Lamarche, M. K. (2013). Measuring organizational evaluation capacity in the Canadian federal government. *Canadian Journal of Program Evaluation*, *28*(2), 1–19. https://doi.org/10.3138/cjpe.28.001

Bradie, M., & Harms, W. (2006). Evolutionary epistemology. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy*. The Metaphysics Research Lab, Center for the Study of Language and Information, Stanford University.

Buckley, J., Hargraves, M., & Moorman, L. (2021). The relational nature of evaluation capacity building: Lessons from facilitated evaluation partnerships. *New Directions for Evaluation*, *169*, 47–64.

Buckley, J., Postlewaite, E., Archibald, T., Linver, M. R., & Urban, J. B. (2024). A protocol for participatory data use. American Journal of Evaluation. Advance online publication. https://doi.org/10.1177/10982140241234835

Campbell, D. T. (1974). Evolutionary epistemology. In P. A. Schilpp (Ed.), *The philosophy of Karl Popper* (pp. 413–463). Open Court Publishing Co.

Campbell, D. T. (1988). Evolutionary epistemology. In E. S. Overman & D. T. Campbell (Eds.), *Methodology and epistemology for social science: Selected papers* (1st ed.). University of Chicago Press.

Cziko, G. A., & Campbell, D. T. (1990). Comprehensive evolutionary epistemology and bibliography. *Journal of Social and Biological Structures*, *13*(1), 41–82. https://doi.org/10.1016/0140-1750(90)90033-3

Danseco, E. (2013). The 5 Cs for innovating in evaluation: Lessons from the field. *Canadian Journal of Program Evaluation*, *28*(2), 107–117.

Darwin, C. (1859). *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. John Murray.

Donaldson, S. I., & Grant-Vallone, E. J. (2002). Understanding self-report bias in organizational behavior research. *Journal of Business and Psychology*, *17*, 245–260. https://doi.org/10.1023/A:1019637632584

Frierson, H. T., Hood, S., Hughes, G. B., & Thomas, V. G. (2010). A guide to conducting culturally responsive evaluations. In J. F. Westat (Ed.), *The 2010 user-friendly handbook for project evaluation* (pp. 75–96). National Science Foundation.

Head, C. B., Jasper, P., McConnachie, M., Raftree, L., & Higdon, G. (2023). Large language model applications for evaluation: Opportunities and ethical implications. *New Directions for Evaluation*, *2023*(178–179), 33–46. https://doi.org/10.1002/ev.20556

Hood, S., Hopson, R., & Frierson, H. (Eds.). (2014). *Continuing the journey to reposition culture and cultural context in evaluation theory and practice*. Information Age Publishing.

Hood, S., Hopson, R. K., & Kirkhart, K. E. (2015). Culturally responsive evaluation. In K. E. Newcomer, H. P. Hatry, J. S. Wholey (Eds.), *Handbook of practical program evaluation* (pp. 281–317). Wiley.

Hopson, R. K. (2003). *Overview of multicultural and culturally competent program evaluation: Issues, challenges and opportunities*. California Endowment.

1534875x, 2024, 183, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ev.20611 by <Shibboleth>-member@8082490.imf.org, Wiley Online Library on [25/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Hopson, R. K. (2009). Reclaiming knowledge at the margins: Culturally responsive evaluation in the current evaluation moment. In K. Ryan & B. Cousins (Eds.), *International handbook of educational evaluation* (pp. 429–446). SAGE Publishing.

House, E. R. (2019). Evaluation with a focus on justice. *New Directions for Evaluation*, *163*, 61–72.

Hurteau, M., & Archibald, T. (Eds.). (in press). *Practical wisdom for an ethical evaluation practice*. Information Age Publishing.

Jubilee Centre for Character and Virtues. (2022). *The Jubilee Centre framework for character education in schools*. Jubilee Centre for Character and Virtue. Retrieved from https://www.jubileecentre.ac.uk/character-education-/the-jubilee-centre-framework-for-character-education-in-schools/

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

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. https://doi.org/10.1177/1098214011434608

Laszlo, E. (1996). *The systems view of the world: A holistic vision for our time*. Hampton Press.

Lerner, R. M. (2006). Developmental science, developmental systems, and contemporary theories of human development. In R. M. Lerner, & W. Damon (Eds.), *Handbook of child psychology: Theoretical models of human development* (6th ed., Vol. 1, pp. 1–17). John Wiley & Sons.

Mayr, E. (2001). *What evolution is*. Basic Books.

Mertens, D. M. (2009). *Transformative research and evaluation*. Guilford Press.

Midgley, G. (2003). *Systems thinking*. SAGE.

O'Brien, S., McNamara, G., O'Hara, J., & Brown, M. (2017). External specialist support for school self-evaluation: Testing a model of support in Irish post-primary schools. *Evaluation*, *23*(1), 61–79. https://doi.org/10.1177/1356389016684248

Overton, W. F. (2006). Developmental psychology: Philosophy, concepts, methodology. In R. M. Lerner & W. Damon (Eds.), *Handbook of child psychology: Theoretical models of human development* (6th ed., Vol. 1, pp. 18–88). John Wiley & Sons.

Overton, W. F. (2010). Life-span development: Concepts and issues. In W. F. Overton & R. M. Lerner (Eds.), *Handbook of life-span development: Cognition, biology, and methods* (Vol. 1, pp. 1–29). Wiley.

Patton, M. Q. (1982). *Practical evaluation*. SAGE Publications.

Patton, M. Q. (2008). *Utilization-focused evaluation*. SAGE.

Popper, K. (1973). *Evolutionary epistemology* (Sections I–VI of the rationality of scientific revolutions). Herbert Spencer lecture. University of Oxford.

Popper, K. R. (1984). Evolutionary epistemology. In D. W. Miller (Ed.), *Popper selections* (pp. 78–86). Princeton University Press.

Preskill, H. (2008). Evaluation's second act: A spotlight on learning. *American Journal of Evaluation*, *29*(2), 127–138.

Ragsdell, G., West, D., & Wilby, J. (Eds.). (2002). *Systems theory and practice in the knowledge age*. Kluwer Academic/Plenum Publishers.

Rogers, A., Radcliffe, D., Babyack, S., & Layton, T. (2018). Demonstrating the value of community development: An inclusive evaluation capacity building approach in a nonprofit Aboriginal and Torres Strait Islander organisation. *Evaluation Journal of Australasia*, *18*(4), 234–255. https://doi.org/10.1177/1035719X18803718

Sarti, A. J., Sutherland, S., Landriault, A., DesRosier, K., Brien, S., & Cardinal, P. (2017). Understanding of evaluation capacity building in practice: A case study of a national medical education organization. *Advances in Medical Education and Practice*, *8*, 761–767. https://doi.org/10.2147/AMEP.S141886

Schwandt, T. (2015). *Evaluation foundations revisited: Cultivating a life of the mind for practice*. Stanford University Press.

Schwandt, T., & Gates, E. (2021). *Evaluating and valuing in social research*. Guilford Publications.

Suarez-Balcazar, Y., & Taylor-Ritzler, T. (2014). Moving from science to practice in evaluation capacity building. *American Journal of Evaluation*, *35*(1), 95–99. https://doi.org/10.1177/1098214013499440

Taylor-Ritzer, T., Suarez-Balcazar, Y., Garcia-Iriarte, E., Henry, D. B., & Balcazar, F. E. (2013). Understanding and measuring evaluation capacity: A model and instrument validation study. *American Journal of Evaluation*, *34*(2), 190–206.

Tilton, Z., LaVelle, J. M., Ford, T., & Montenegro, M. (2023). Artificial intelligence and the future of evaluation education: Possibilities and prototypes. *New Directions for Evaluation*, *2023*(178–179), 97–109.

Trochim, W., Urban, J. B., Hargraves, M., Hebbard, C., Buckley, J., Archibald, T., Johnson, M., & Burgermaster, M. (2016). *The guide to the systems evaluation protocol* (V 3.1). Cornell Digital Print Services.

Urban, J. B., Archibald, T., Hargraves, M., Buckley, J., Hebbard, C., Linver, M. R., & Trochim, W. M. (2021). Introduction to relational systems evaluation. *New Directions for Evaluation*, *2021*(169), 11–18.

Urban, J. B., Hargraves, M., Buckley, J., Archibald, T., Hebbard, C., & Trochim, W. M. (2021). The systems evaluation protocol for evaluation planning. *New Directions for Evaluation*, *2021*(169), 31–45.

1534875x, 2024, 183, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ev.20611 by <Shibboleth>-member@80182490.imf.org, Wiley Online Library on [25/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

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. https://doi.org/10.1177/1558689813516388

Urban, J. B., Hargraves, M., & Trochim, W. M. (2014). Evolutionary evaluation: Implications for evaluators, researchers, practitioners, funders and the evidence-based program mandate. *Evaluation and Program Planning*, *45*, 127–139. https://doi.org/10.1016/j.evalprogplan.2014.03.011

Urban, J. B., Linver, M. R., Buckley, J., Hargraves, M., & Archibald, T. (2024). Character-minded evaluation: Recognizing and activating the essential role of character in high-quality evaluation. In M. D. Matthews & R. M. Lerner (Eds.), *Routledge international handbooks of multidisciplinary perspectives on character development. Volume 1. Conceptualizing and defining character* (pp. 310–334). Routledge.

Urban, J. B., Linver, M. R., Johnson, S. K., MacDonnell, M., Chauveron, L., Glina, M., & Gama, L. (2018). Developing the next generation of engaged youth: Inspire-Aspire Global Citizens in the making. *Journal of Moral Education*, *47*(1), 104–125. https://doi.org/10.1080/03057240.2017.1396967

Wingate, L. A., Robertson, K., Fitzgerald, M., Rucks, L., Tsuzaki, T., Clasen, C., & Schwob, J. (2022). Thinking outside the self-report: Using evaluation plans to assess evaluation capacity building. *American Journal of Evaluation*, *43*(4), 515–538. https://doi.org/10.1177/10982140211062884

How to cite this article: Urban, J. B., Postlewaite, E. L., Davis, B., Les, E., Buckley, J., & Hargraves, M. (2024). Relational Systems Evaluation rubrics as tools for building and measuring evaluation capacity. *New Directions for Evaluation*, *2024*, 59–79. https://doi.org/10.1002/ev.20611

Appendix A
Relational systems evaluation (RSE) evaluation plan rubric

Reprinted with permission from the authors

This rubric is intended to be a tool for reviewers to use in providing systematic scores on evaluation plans for programs. There are items or sections in this rubric that may not be applicable in certain situations. For example, not all evaluation plans include a programmatic visual theory of change model. Therefore, the first step in using this rubric is to remove the items/sections that are not applicable to the situation or for which the necessary information is not available to the reviewer.

Directions

1. Read through the entire evaluation plan being reviewed.
2. Complete the scoring by section.
3. Score every item (omitting any that have been designated or agreed upon as not applicable to your situation).
4. Items absent from the evaluation plan should be scored as 0 = *Absent/missing*
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;

1534875x, 2024, 183, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ev.20611 by <Shibboleth>-member@80182490.imf.org, Wiley Online Library on [25/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

- 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 (i.e., the elements of the evaluation plan—the evaluation purpose, questions, measures, sampling strategy, design, and analysis plans—should be consistent with each other).

Section-by-section assessment

The categories below correspond to typical 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. Response format is a 0–4 scale, where 0 = *Absent/missing*, 1 = *Minimally acceptable*, 2 = *Adequate*, 3 = *Good*, and 4 = *Excellent*. The highest possible score is 196. For each item, please refer to the evaluation plan (and visual theory of change model [i.e., logic model] as necessary).

Evaluation plan elements

Program description

1. *Communication of goals*: The plan clearly conveys the major goals of the program being evaluated.
2. *Program implementation*: There is a description of the program implementation (e.g., program scale, activities).
3. *Target audience*: The key features of the social, political, and cultural background of the participants and individuals the program is intended to impact are described.
4. *Program context*: The description includes the key features of the program context (e.g., includes information about the social, cultural, and physical context in which the program takes place).
5. *Program background*: The history of the program's development is described, and/or the description refers to the research evidence that the program draws upon or is related to.

Evaluation purpose

6. *Program activities/outcomes/assumptions*: The specific program activities, outcomes, or assumptions that are the focus of evaluation are identified.
7. *Main goals*: The main goal(s) of the evaluation are articulated, including how the evaluation will serve communities affected by the evaluation or the program.
8. *Intended uses*: There is a description of the intended use(s) of the evaluation, including how it will be useful to the organization, to participating or affected communities, or to other key stakeholders.
9. *Fits with other evaluations*: There is an explanation of how the current evaluation plan fits in with any other (prior or ongoing) evaluation work on this program.
10. *Appropriateness*: The evaluation goals are appropriate relative to the stage of development of the program (e.g., Initiation phase [newly developed or adapted], Development phase [still being adjusted after prior rounds]; Stability phase [steady implementation with limited changes]; Dissemination phase [being implemented in multiple sites]) and its prior evaluation(s).

Evaluation Questions (*Note*: Some evaluation plans may use other terms [e.g., objectives, goals, hypotheses, aims] instead of "evaluation questions.")

11. *Alignment*: The evaluation questions are explicitly derived from and fulfill the stated evaluation purpose.

1534875x, 2024, 183, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ev.20611 by <Shibboleth>-member@80182490.imf.org, Wiley Online Library on [25/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

12. *Clarity*: The evaluation questions are stated clearly and precisely enough to serve as a guide for the evaluation plan.
13. *Logic model alignment*: There is alignment between the evaluation questions and the program's logic (e.g., questions are clearly related to the program's theory of change).
14. *Appropriateness*: The evaluation questions are appropriate given the stage of development of the program (e.g., Initiation phase [newly developed or adapted], Development phase [still being adjusted after prior rounds]; Stability phase [steady implementation with limited changes]; Dissemination phase [being implemented in multiple sites]) and its prior evaluation(s).
15. *Feasibility*: The number of questions appears manageable.

Design

16. *Alignment*: There is a design to address each/all evaluation question/s.
17. *Description of design*: There is sufficient detail describing the design (e.g., post-only, pre/post, pre/post with comparison group, case study, nonequivalent groups, time series, longitudinal, cross-sectional, sequential mixed-method) to provide clarity about what is intended and when. (*Note*: Just naming the design as quantitative, qualitative, or mixed method is not sufficient.)
18. *Generation of evidence*: The selected design(s) is appropriate for generating valid evidence to answer evaluation question(s).
19. *Appropriateness*: The selected design(s) is appropriate given the stage of development of the program (e.g., Initiation phase [newly developed or adapted], Development phase [still being adjusted after prior rounds]; Stability phase [steady implementation with limited changes]; Dissemination phase [being implemented in multiple sites]) and its prior evaluation(s).
20. *Burden*: The overall design is appropriate given the organization's capacity, feasibility of implementation, and burden on respondents.

Measures

21. *Alignment of measure type*: There is a measure type(s) identified for each evaluation question (e.g., survey, observation, interview).
22. *Measure description*: Each measure is described in sufficient detail to determine measure appropriateness (including specifying the focal construct).
23. *Appropriateness of measure*: The type of measure(s) selected (e.g., survey, observation, interview) is appropriate for generating valid evidence to answer evaluation question(s).
24. *Measure fit*: The measure(s) have been reviewed, vetted, and/or tested by/with members of the target population (for program setting, population, audience, etc.), or there is a plan to do so.
25. *Measure origin*: There is a description of the origin of each measure (e.g., appropriately cited, development of new tool described, including any use of AI).
26. *Cultural responsiveness*: There is a plan to offer culturally responsive support to assist with collecting data from persons whose participation would otherwise be limited by language, abilities, or factors such as familiarity or trust.

Sampling

27. *Alignment*: There is a sampling strategy/technique for each evaluation question.

1534875x, 2024, 183, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ev.20611 by <Shibboleth>-member@8018249.imf.org, Wiley Online Library on [25/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

28. *Population*: There is a description of the population(s) of interest from which the sample(s) is drawn (e.g., size, scope/scale).
29. *Sample*: There is a description of the sample(s).
30. *Recruitment*: There is a description of the recruitment strategy for the sample(s) that is sensitive to the needs and conditions of the target population.
31. *Sampling technique*: The choice of sampling technique(s) (e.g., techniques such as simple random, convenience, cluster random) will generate appropriate evidence.
32. *Sample size*: The sample size is appropriate for generating sufficient evidence to address the evaluation question(s).

Data collection and management

33. *Alignment*: There is a plan for how data will be collected logistically (e.g., in person, electronically) for each measure.
34. *Data storage and prep*: There is a complete and appropriate plan for how the data will be organized and stored securely in preparation for analysis for each measure.
35. *Back-up plans*: There are backup plans in place for unexpected circumstances (e.g., technology changes, global pandemic, natural disasters).

Data analysis

36. *Alignment*: There is a data analysis strategy for each evaluation question.
37. *Appropriateness*: The data analysis strategies are appropriate for generating valid evidence to answer evaluation question(s).

Evaluation reporting and utilization

38. *Purpose alignment*: There is an evaluation reporting/utilization strategy for each intended use stated in the evaluation purpose section.
39. *Comprehensiveness*: The reporting plan addresses external and internal reporting, including the form and frequency of reporting.
40. *Culturally appropriate*: The reporting plan is culturally appropriate, including customizing reporting methods for stakeholder audiences that address communication and language issues as needed. This may involve adapting/creating multiple reports and utilizing various reporting methods.
41. *Report use*: The reporting plan(s) is appropriate for the utilization of the evaluation results given the evaluation purpose.
42. *Appropriateness*: The reporting plan(s) is appropriate for the utilization of the evaluation results relative to the program's current stage of development (e.g., Initiation phase [newly developed or adapted], Development phase [still being adjusted after prior rounds]; Stability phase [steady implementation with limited changes]; Dissemination phase [being implemented in multiple sites]).

Evaluation timeline

43. *Program activities*: The timeline includes key program activities and/or events relevant to the evaluation design (e.g., events around which a pre/post-test will be administered).

1534875x, 2024, 183, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ev.20611 by <Shibboleth>-member@80182490.imf.org, Wiley Online Library on [25/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

44. *Evaluation activities*: The timeline includes the key evaluation activities (e.g., measure development/procurement, data collection, analysis).
45. *Specificity*: The timeline events are given in calendar time, not just in relative terms.

Overall

46. *Writing*: The quality of writing (e.g., clarity, consistency of voice, correct grammar, and proper spelling) is high.
47. *Communication tool*: The evaluation plan is an effective communication tool (e.g., language and phrasing are understandable to outside readers).
48. *Alignment*: There is internal alignment throughout the evaluation plan (i.e., sample, design, measurement, and analysis plans are consistent with the evaluation questions and with each other).
49. *Feasibility of the evaluation plan*: The timeline, budget, personnel, and resources required to complete the evaluation seem reasonable given the scope of the plan.

Appendix B
Relational systems evaluation (RSE) visual theory of change rubric
Reprinted with permission from the authors

Visual depictions of a theory of change can take many forms. Relational systems evaluation promotes the use of pathway models, which are a type of logic model or theory of change. See Urban and colleagues' (Urban, Hargraves, Buckley, et al., 2021) manuscript for more information about pathway models as part of the Systems Evaluation Protocol. The key features of the theory of change model that are fundamental to its definition are the nodes/boxes that name program activities and anticipated outcomes, and the arrows that represent the chain of logic that connects these nodes/boxes. In some instances, people use visual theory of change models for something other than a theory of change. If that is the case, the user will need to adapt the rubric.

In this rubric, an individual theory of change model is viewed as having been developed for a "purpose" (e.g., communication, accountability, program planning), and these vary depending on the intended audience. Since this purpose is typically not known by the person scoring the rubric, the rubric is designed to assess the internal coherence and technical quality of a theory of change model but not whether the theory of change model is aligned with its purpose. Similarly, this rubric cannot be used to score alignment between the model and the program's context or assumptions (e.g., important characteristics of participant groups or community), since that information is also not known to the scorer. This rubric is designed to be scored by a person with at least a basic understanding of evaluation (e.g., by a funder or an internal team for capacity building).

Rubric scoring

The categories below correspond to the theory of change model elements. Each item within a category provides a specific criterion for quality of work. The scale is intended for numerical scoring. The rubric consists of four scorable sections. Several items are scored with a binary 0 or 1, where 0 = *Missing* and 1 = *Present*. The other items are scored using a 0–4 scale, where 0 = *Unacceptable*, 1 = *Needs significant improvement*, 2 = *Minimally acceptable*, 3 = *Good*, and 4 = *Excellent*. Scores should be summed to provide an overall score that represents the quality of the theory of change model. The highest possible score is 33. Higher scores represent higher-quality theory of change models. Lower scores represent lower-quality theory of change models.

1534875x, 2024, 183, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ev.20611 by <Shibboleth>-member@80182490.imf.org, Wiley Online Library on [25/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

1534875x, 2024, 183, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ev.20611 by <Shibboleth>-member@80182490.imf.org, Wiley Online Library on [25/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Visual theory of change model elements

Conduct a visual inspection of the model's structure, including labels and legends. Do not worry about the text inside the boxes. (*Scoring note*: Items 1–4 should receive a binary score of 1 or 0.)

1. *Dead ends*: Pathways go all the way to long-term outcomes (i.e., there are no dead ends). (Score 0–1)
2. *Outcomes*: All outcomes can be traced back to an activity (i.e., there are no "miracle outcomes"). (Score 0–1)
3. *Sequencing*: The sequencing and progression of outcomes follows from short-term to mid-term to long-term. Examine the model to ensure that (a) no short-term outcome directly connects to a long-term outcome; (b) no activity directly connects to a mid-term outcome; (c) no activity directly connects to a long-term outcome; (d) no activity is connected to another activity with an arrow. (Score 0–1)
4. *Outputs*: Any outputs included in the model are terminal nodes (i.e., there are no arrows connecting outputs to outcomes). Outputs have an arrow(s) coming from an activity(ies) but no arrow going out to anything else. (Score 0–1)

Nodes (Boxes):

Read and score what is inside the boxes. (*Scoring note*: Items 5–7 should be scored on a scale of 0–4. Item 8 should receive a binary score of 1 or 0.)

5. *Interpretability*: The theory of change model is interpretable by outside viewers or external stakeholders. In other words, the text inside the boxes (activities and outcomes) is readable and understandable and there is no incomplete information. (Score 0–4)
6. *Activities*: Activities are the elements of the program that participants experience or interact with; they do not include change language. Some examples of activities include workshops, outreach campaigns, and access to mentors. (Score 0–4)
7. *Outcomes*: Outcomes are phrased as the addition or emergence of something that did not previously exist and is new (e.g., network, attitude, behavior, knowledge) or that changed (e.g., increased, improved, decreased); outcomes include the who (what person or group is experiencing the change), what (what is changing), and how (the nature of change, if appropriate). For example, the outcome "Youth interest in science increases" is more interpretable than the outcome "Interest increases." Another example: The outcome "A new youth network is formed" is more interpretable than the outcome "Youth network." Outcomes are not double-barreled; each one identifies a single variable. (Score 0–4)
8. *Outputs*: Outputs are phrased as tangible byproducts of an activity. (Score 0–1; if there are no outputs in the model, assign a score of 1.)

Links (Arrows):

Read the contents of the nodes (boxes), and evaluate the connections between them. (*Scoring note*: Items 9–11 should be scored on a scale of 0–4.)

9. *Leaps in logic*: Every arrow represents a causal or attributive assumption. No arrow(s) represent(s) a larger leap relative to the other arrows in the model. Large leaps may signal that there are important or plausible nodes not represented in the model. (Score 0–4)
10. *Missing links*: Obvious/intuitive links among nodes (boxes) are not missing from the model. Potential signals there are important or plausible connections that are not

represented in the model: (a) Pathways are independent of one another (there is no crossover among pathways); or (b) there is only one node for each type of outcome (e.g., one short-term outcome, one mid-term outcome, one long-term outcome) in a pathway. (Score 0–4)

11. *Too many links*: Redundancy of links is avoided. Only the arrows that represent the important linkages in the chain of logic are included, in order to reduce visual clutter. (Score 0–4)

Overall quality:

As a scorer, define what you mean by this. (*Scoring note*: This item is scored holistically on a scale from 0–4.)

12. *Overall quality*: One way to define overall quality is to assess whether the theory of change model aligns with the program description or the theory of change model purpose. The theory of change model should provide insights into how the program works and how change unfolds. In other words, it should be informative. The level of detail should be within a useful range (i.e., not too small and not too large). (Score 0–4)

AUTHOR BIOGRAPHIES

Jennifer Brown Urban is a professor in the Department of Family Science and Human Development and co-director of the Institute for Research on Youth Thriving and Evaluation, both at Montclair State University.

Elyse L. Postlewaite is a postdoctoral researcher at the Institute for Research on Youth Thriving and Evaluation at Montclair State University.

Bekki Davis is an assistant director of student services at Montclair State University and a PhD student in Montclair State University's Department of Family Science and Human Development.

Elaine Les is a PhD student at Montclair State University in the Department of Family Science and Human Development and a doctoral research fellow at the Institute for Research on Youth Thriving and Evaluation.

Jane Buckley is an independent consultant who specializes in evaluation capacity building, evaluative thinking, and character-minded evaluation.

Monica Hargraves has been involved in research on evaluation and evaluation capacity building since 2008 when she joined the Cornell Office for Research on Evaluation; she continues this work now as an independent consultant.

1534875x, 2024, 183, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ev.20611 by <Shibboleth>-member@80182490.imf.org, Wiley Online Library on [25/11/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License