The Foundation Review

a publication of the Dorothy A. Johnson Center for Philanthropy at Grand Valley State University

Volume 16
Issue 2 *Democracy, Equity, and Power*

10-2024

Raising the Bar: Improving How to Assess Evidence Quality in Evaluating Systems-Change Efforts

Marina Apgar
Institute of Development Studies

Thomas Aston
independent consultant

Mieke Snijder
Institute of Development Studies

Tom Zwollo
Save the Children Netherlands

Follow this and additional works at: https://scholarworks.gvsu.edu/tfr

 Part of the Nonprofit Administration and Management Commons, Public Administration Commons, Public Affairs Commons, and the Public Policy Commons

Recommended Citation

Apgar, M., Aston, T., Snijder, M., & Zwollo, T. (2024). Raising the Bar: Improving How to Assess Evidence Quality in Evaluating Systems-Change Efforts. *The Foundation Review, 16*(2). https://doi.org/10.9707/1944-5660.1712

Copyright © Dorothy A. Johnson Center for Philanthropy at Grand Valley State University. The Foundation Review is reproduced electronically by ScholarWorks@GVSU. https://scholarworks.gvsu.edu/tfr

doi: 10.9707/1944-5660.1712

Raising the Bar: Improving How to Assess Evidence Quality in Evaluating Systems-Change Efforts

Marina Apgar, Ph.D., Institute of Development Studies; Thomas Aston, Ph.D., Independent Consultant; Mieke Snijder, Ph.D., Institute of Development Studies, and Tom Zwollo, M.Sc., Save the Children Netherlands

Keywords: *Values, rigor, participation, evidence, rubrics*

Introduction

There is a long-standing debate regarding what counts as rigorous and credible evidence for evaluation (Donaldson et al., 2008; Mosley et al., 2024). Yet, there is less discussion on how best to assess rigorous evidence related to complex programming contexts, and what might constitute relevant criteria for such an assessment (Preskill & Lynn, 2016; Schwandt & Gates, 2021; Aston et al., 2021; Aston & Apgar, 2022). Rigor has often been reduced to a discussion of evidence hierarchies, usually focused on the supposed "gold standard" of randomized control trials and the "what works" agenda, couched within evidence clearing houses (Boruch & Turner, 2023). As Howard White (2019) explains, this agenda has dominated what "counts" as valid knowledge and rigorous evidence, fusing assessment of evaluation methods with assessment of evidence.

However, evidence hierarchies have been critiqued as misleading (Nutley et al., 2013). Randomized control trials (RCTs) are not always appropriate, feasible, or even ethical (Befani et al., 2015; Schwandt & Gates, 2021). They are designed with assumptions about control, stability, and fidelity which rarely hold in complex intervention contexts or at scale. It is argued, therefore, that RCTs are inappropriate to assess systems change (Bicket et al., 2020; Lynn et al., 2021). On the other hand, there have been several efforts to debunk myths about the supposed lack of rigor of nonexperimental evaluation approaches (Lynn et al., 2021; Raimondo, 2023). Lynn and colleagues demonstrate that experimental and quasi-experimental methods are not the only ways to assess cause-and-effect relationships and argue that philanthropy needs

Key Points

- Facing the great scale of societal challenges, philanthropic organizations are increasingly calling for systems change. Evaluating systems change requires innovative approaches that respond to the complexities of such change in ways that support equity and multiracial democracy rather than undermining them.
- A key concern in evaluating systems change is how to do so rigorously. Rigor has traditionally been equated with evaluative criteria such as independence and objectivity, and experimental methods and evidence hierarchies which sit uncomfortably with both complexity and equity. Yet when taking an alternative approach, many philanthropic organizations fear that without these standards, there are no standards at all.
- Establishing means to assess evidence standards is a key challenge for complexity-informed evaluation. This article argues that more appropriate, flexible, and inclusive standards for assessing evidence quality in systems-change efforts are achievable. Based on a review of evidence standards, learning from the causal pathways and inclusive rigor networks, and using the example of evaluation of the CLARISSA program, it lays out a set of principles and tools to guide assessment by philanthropic organizations of evidence quality in systems-change evaluation.

to examine causal relationships through a growing suite of methodological approaches as relevant to different systems-change strategies.[1] Despite these trends, philanthropic evaluation tends to still rely on descriptive measurement and analysis, such as the performance measurement approach recently proposed by Brown and Rosser (2023).

Lynn and Coffman (2024) usefully distinguish two mental models of systems change: system emergence and system dynamics. In the first, strategies informed by complexity theory assume that it is impossible to predict the type of change that might emerge in a system, requiring evaluation to look back once change has emerged to retrospectively explore and learn from causal pathways, considering relevant factors that together have created change in a system. In these conditions, most traditional pre- and post-evaluation designs would be inappropriate. Approaches that map system dynamics and identify leverage points for strategic interventions, on the other hand, may predefine some system domains as the focus of evaluation while still being open to dynamic interactions in the system. Approaches that focus on discrete parts of the system can assume greater predictability and could be served by evaluation approaches that theorize the intended pathways at the outset and empirically test if and how change unfolds. Both approaches call for the use of causal methodologies that open up the "black box" of systems-change strategies. Some philanthropic organizations argue that communities themselves also need to play a role in explaining these strategies (Carr & Morariu, 2023). For this, philanthropy requires a more inclusive understanding of rigor that gives space for plural perspectives to inform more useful evaluation approaches.

In this article, we build on thinking emerging from several communities of practice considering alternative approaches to rigor to show how more appropriate, flexible, and inclusive standards for assessing evidence quality in systems-change efforts are achievable. The Inclusive Rigor Co-Lab, funded by Humanity United, built on the work of Robert Chambers (2015), who reframed rigor as inclusive to embrace complexity. His seven canons (eclectic methodological pluralism; improvisation and innovation; adaptive iteration; triangulation; plural perspectives; optimal ignorance; and appropriate imprecision — being open, alert, and inquisitive) stem from a participatory epistemology that underpins inclusive methods. From this perspective we should be attentive to context, appreciate participatory and iterative evaluation co-design, and consider the roles of evaluators as reflexive facilitators rather than objective judgers (Aston et al., 2021; Apgar et al., 2024a).

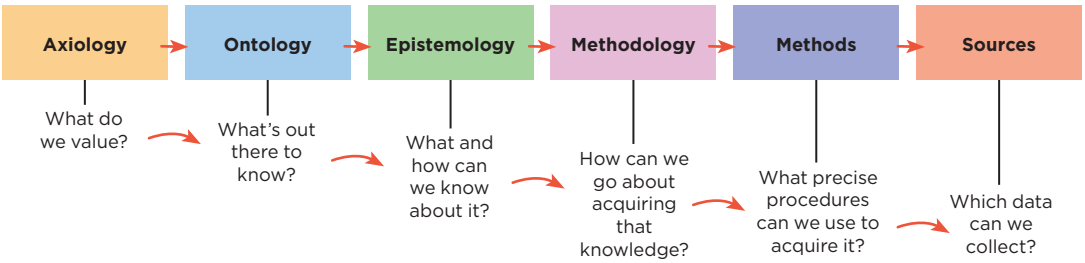
Notions of "adaptive rigor" have also recently evolved in response to complexity-aware evaluation that aims to harness actionable learning (Preskill & Lynn, 2016; Wild & Buffardi, 2019; Aston & Apgar, 2022). The Causal Pathways Initiative, launched by the Walton Family Foundation, builds on both "inclusive" and "adaptive" approaches to rigor as it seeks to support philanthropy to build awareness, will, and skills to use evaluation approaches that can make sense of causal relationships while paying attention to equity. Among the complementary aims of these communities is understanding what "good" looks like for more complexity-aware approaches, and how to raise the bar. Before assessing what is "good," however, it is necessary to re-center evaluation in values as the starting point for redefining rigor.

Values and Evidence Standards

Values are making a comeback in evaluation. Thomas Schwandt and Emily Gates (2021) define valuing as a "kind of practice that involves identifying, naming, considering, and holding or respecting something … as important, beneficial, right to do, good to be" (p. vii); and they define evaluating as a "particular kind of empirical investigation … appraising, weighing up, assessing, calculating, gauging, rating, and ranking" (p. vii). At the heart of both valuing and evaluation is criteria — principles or standards

[1] See Lynn & Apgar (2024), which explores several approaches to this.

FIGURE 1 Sequencing Questions to Define Rigor in Ways that Align with Underpinning Values



that different stakeholders value. Before assessing or rating, we must first establish what we value — most. Evaluators and foundations need to make choices about what they value and critically engage with the socially constructed notions of meaning in different contexts.

As Brown and Dueñas (2020) illustrate, we first need to understand what we value (axiology) before addressing what there is to know (ontology), how we can know about it (epistemology), and how to collect evidence to better understand what we aim to know better (methodology, methods, sources). (See Figure 1.) The Equitable Evaluation Framework™ espouses a similar axiological and epistemic perspective related to how values inform what can be known and what counts as rigor (Chilisa, 2019; Lowther & McKegg, 2023; Coné & Dean-Coffey, 2024).

There are several applications of this thinking. Gates and colleagues (2024), for example, provide a framework that supports explicit criteria specification, based on a combination of deliberative democratic and critical approaches that focus on the need to deliberate between plural values, while navigating power to ensure inclusion.

Our starting proposition, therefore, is that values are the basis upon which foundations can establish the criteria that matter most, rather than simply using existing criteria without critically examining what underpins them. The Campbellian validity framework of statistical conclusion, internal, construct, and external validity, for example, was based on validity criteria appropriate for quantitative methods, yet remains dominant today and often is applied generically (Lund, 2021). A review by Downes and Gullickson (2022) on what "valid" means in evaluation found 40 different conceptualizations in use, showing that validity is more contested and multifaceted than assumed. If foundations champion equity or community participation, for example, and make these values explicit, then there are several relevant quality criteria they may wish to consider, such as multicultural validity, responsiveness, and transferability (Kirkhart, 2010; Aston et al., 2021).

The turn to values is aligned with a call for evaluation to be geared toward questions and criteria, rather than driven by particular method or data preferences (Stern et al. 2012; Gates et al., 2024). Schwandt and Gates (2021) point out that "choosing criteria commits the evaluator to look for certain kinds of evidence and to appeal to certain kinds of warrants … to justify resulting evaluative claims" (p. 2). A central point here is that evidence is not good or bad a priori, but rather depends on what that evidence is supposed to prove — i.e., its potential probative value (Schwandt, 2008). This should be defined by the users themselves, rather than be driven by methodological choice alone. In the context of evaluation that centers equity, the starting point, we argue, must be an expanded view of users, which invites us to first consider the question of whose values count (Chambers, 2015).

What Values Matter to Whom?

We have had numerous discussions about criteria used to assess the quality of evidence with a range of evaluation practitioners, researchers, commissioners, and programmers from diverse

TABLE 1 Synthesis of Criteria Commonly Used to Judge the Quality of Evidence

Ranking	Training on Contribution Analysis	Training on Assessing Strength of Evidence	Frequency	Symposium with U.S. Philanthropic Audience
Highest	Transparency	Utilization	Highest	Credibility/ Triangulation/ Utilization
	Triangulation	Transparency		Participatory
	Replicability	Independence/ Triangulation/ Uniqueness		Equitability
	Reliability	Responsiveness/ Transferability/ Ethics		Reliability
Lowest	Utilization	Plausibility	Lowest	Novelty

contexts, through a series of trainings we deliver and work within our own communities of practice (Centre for Development Impact, Inclusive Rigour Co-Lab, and Causal Pathways Initiative). In this section, we reflect across these conversations to shed light on the question of which criteria matter to whom, illustrating the diverse entry points different producers and users of evidence might have, and how that then defines what criteria might be appropriate for any given evaluation.

A plethora of evidence assessment frameworks and critical appraisal tools are used by government departments, universities, think tanks, and research and evaluation consultancy firms, yet rarely by foundations (e.g., Puttick & Ludlow, 2013; Specialist Unit for Review Evidence, 2018). Across these we find similar criteria used for the evidence produced by nonexperimental methods (see Aston & Apgar, 2023): transparency, triangulation, ethics, plausibility, uniqueness, independence, responsiveness, and transferability. We have taken these common criteria as a starting point for audiences in different sessions to gauge which criteria they use explicitly or implicitly to judge the quality of evidence. (See Table 1.)

The first two engagements (columns 1 and 2) were attended by 30 and 20 participants respectively, from a largely U.K. evaluation audience. They were asked to rank which criteria they felt were most important in their work. Most participants held research or evaluation roles within U.K. government ministries, while others worked within academic institutions and evaluation consultancy firms; and in each iteration, only two participants worked at philanthropic foundations. The third engagement was during an online session on how to select methods with participants either working in or with U.S. philanthropic evaluation, with 30 participants who were asked to share the criteria they are using. (Therefore, the data from this engagement is about frequency of use rather than a ranking of importance.)

Across the three engagements triangulation is a common criterion, and transparency was ranked as high by the two engagements with U.K. evaluation audiences. It is perhaps unsurprising that these two criteria, which are more easily understood and widely used in qualitative research assessments, are more commonly valued. We see different levels of valuing of utilization, which was ranked as lowest by the first group of participants and highest by the second two. This might be explained by the context of the first group being a training on a particular theory-based evaluation approach and participants

holding research and government roles, where we might expect the focus to be more on quality within the methodological approach being discussed rather than use.

However, other criteria, such as credibility, are more multifaceted and contested (Donaldson et al., 2008). Credibility, which was named more by the third group, is more open to distinct interpretations. Often, part of the interpretation is the idea of finding evidence that links the intervention (or set of interventions) to an outcome in a specific way — otherwise referred to as uniqueness, which was ranked at the same level as triangulation by the second group. Another multifaceted criterion often considered part of credibility is plausibility; this was mentioned only by the second group and was ranked as the lowest.

What is most striking about the different ways in which criteria were valued among these three groups is the explicit mention of participatory and equitability by the third group. This aligns with the focus on equity in philanthropic strategy, and, in particular, in the U.S. Further, valuing independence and replicability by the two first groups reflects British government policy perspectives. We also find that the emphasis from these groups is more on issues related to internal rather than external validity (i.e., transferability). One likely reason for this is that external validity is particularly challenging in the context of complexity.

Our findings from across these conversations underscore the need to initiate a process of defining evidence quality in order to support evaluative judgements through first surfacing values that might otherwise remain hidden. Depending on the specific evaluation use and the diversity of users involved, the right set of criteria to define "good" in this context could differ significantly.

Using Rubrics to Navigate Complexity and Systems Change

While foundations need the flexibility to choose what they value most, they also need some degree of structure to build confidence in how to assess evidence systematically. Checklists are often used to appraise quality, based on the presence or absence of particular characteristics, but the binary categories they create are often too restrictive. In our experience, we have found that rubrics offer a more satisfactory alternative, particularly for complex change processes where the boundaries are fuzzy and where discussion about the boundaries of different criteria, levels, and descriptions is seen as beneficial by evaluation stakeholders. Rubrics are a form of qualitative scale that include the following:

- *criteria,* the aspects of quality or performance of interest (e.g., credibility);
- *standards,* the level of performance or quality for each criterion (e.g., poor/adequate/good); and
- *descriptors,* descriptions or examples of what each standard looks like for each criterion (Green, 2019).

Which criteria, and how many criteria one ought to choose, depends on evaluation purposes expressed by different stakeholders. While criteria such as triangulation, for example, may seem to have a uniform definition, as rubrics are multifaceted, different stakeholders may prefer to focus on different types of triangulation (e.g., data, source, method). Rubrics entail levels of performance or quality for each criterion chosen (e.g., poor/adequate/good). There is no right answer on how many levels are appropriate under all circumstances. However, there are certain rules of thumb for developing rubrics in general which also apply to evidence rubrics (i.e., adding levels only where distinctions are meaningful).

Ultimately, rubrics are a means to determine "what matters rather than what is easy to measure" (Haldrup, 2023, para. 8). They provide an architecture for a deliberative process to discuss, debate, and define what success looks like (King, 2023). Rubrics are increasingly seen to offer an alternative to understanding the multiplicity of factors that make up systems change (Loveridge, 2023). Deliberation is important for assessing

evidence of systems change because such change cannot be predefined, and consequently neither can the specific (usually qualitative) evidence that allows for a nuanced causal explanation of how that change came about.

To illustrate how to use rubrics to assess evidence quality when evaluating a systems-change initiative, we present a case study of the Child Labour Action-Research-Innovation in South and South-Eastern Asia — CLARISSA — program, with which two of the authors have been involved.

Case Study in Using Strength of Evidence Rubrics

The CLARISSA program was a five-year systemic action research program focused on the worst forms of child labor. It was funded by the U.K. Foreign and Commonwealth Development Office, led by the Institute of Development Studies, and implemented through a consortium of international partners including Terre des Hommes, Child Hope, the Consortium for Street Children, and in-country partners in Nepal and Bangladesh.

The starting assumption of the program was that children end up as laborers because of many and often hidden interactions between multiple actors and multiple factors within households, communities, and labor systems. These complex dynamics lead to unpredictable outcomes for children and other sector stakeholders. Knowing when and how to intervene requires a systemic approach to uncover hidden dynamics and identify leverage points for action, yet most interventions continue to focus on predefined solutions of protection and rescue alone or on specific thematic responses such as education instead of work, and, critically, do not include the lived experience of children and other system actors.

The CLARISSA program responded through adopting systemic action research (Burns, 2007) as an implementation modality. The method is a form of participatory action research that aims to understand and intervene in the underlying system dynamics that lead to patterns of exclusion and exploitation of marginalized groups. It is informed by complexity theory and posits that when the system actors themselves make sense of their own experiences and build their own systemic understanding, they become motivated to identify leverage points for action and, as a result, take more effective actions. It is systemic in two ways: (1) it starts from developing an understanding of the causal dynamics that drive system behaviors, and (2) it works with multiple actors across the system in participatory ways.

The Programmatic Approach to Evaluation and Evidence

Given the complexity of child labor, the learning orientation of the program, and the value placed on lived experience and agency of stakeholders to explore and define their own pathways to systems change, evaluation in CLARISSA was not concerned with measuring predefined indicators. Rather, it was designed to understand and analyze causal pathways. The causal pathways were expected to emerge from three levels of engagement:

- micro level, with system actors on specific issues through action research;
- meso level, through influence on dynamics in the supply chains; and
- macro level, through potential shifts in how others in the child labor programming system responded to the systemic evidence CLARISSA would produce and use.

Contribution analysis (Mayne, 2008) was chosen as an overarching approach for its ability to provide both structure and flexibility in how causal theories of change are nested and explored at multiple levels of engagement. It emphasizes the iterative use of causal theories of change as the program evolves and adapts and acknowledges multiple perspectives as central to the causal analysis required for the exploration of potential pathways, as well as retrospective discovery of how pathways actually took shape (Apgar et al.,

2020). The program's modular evaluation design identified several causal hotspots,[2] and combinations of appropriate methods were selected to respond to each.[3]

The funder was involved in lengthy discussions on the program's overarching approach to evaluation, in particular during the inception period as the partnership was solidifying and the program was taking shape on the ground. During this initial period, differences in assumptions held by partners around what counts as a "rigorous" evaluation design were surfaced, creating some tensions. The evaluation team worked with the program management team to facilitate debate among partners and the funder on these tensions. This led to agreement on the appropriateness of contribution analysis. Given these different starting positions, the evaluation team made explicit the program's approach to evidence as plural in the MEL framework — valuing and using multiple forms, including lived experience and practitioner learning alongside formal research evidence (CLARISSA, 2018). This plural approach, which was agreed to by the funder, and the mix of methods used meant that there were no predefined criteria to define the quality of evidence for the program. Given that what counts as rigorous and credible evidence is contested (Donaldson et al., 2008), the program recognized a multitude of possible criteria could be used.

The evaluation team made its quality criteria explicit, and developed a set of evidence rubrics that could be applied throughout the evaluation as evidence was gathered on emergent pathways to systems change. The team facilitated a deliberative process working across program stakeholders, including in-country CLARISSA staff (facilitating the participatory interventions) and the thematic research team (building evidence on child labor through participatory and qualitative research). At this stage, the funder was not involved in detailed deliberations, having agreed to the broad approach. The evaluation team initiated the process by reviewing all possible criteria based on Downes & Gullickson (2022) and Aston & Apgar (2022), and proposed a set of criteria to the program team. In this first proposal, the team excluded "independence" and "generalizability" as inappropriate, given that the evaluation was to be conducted internally and aimed to provide nuanced responses to causal questions, paying particular attention to how processes worked in context.

Evaluation and thematic research teams deliberated on what criteria were appropriate for all forms of evidence emerging from the program, and where distinct criteria were needed for making causal inferences (evaluation research). Three core criteria were agreed across all forms of evidence produced by the program:

- *Transparency.* Given that most of CLARISSA's evaluation and research methods were qualitative and focused on uncovering hidden dynamics in supply chains and systems, making explicit the processes through which data were collected and analysis was undertaken, and by whom, was a foundational criterion.

- *Representativeness.* This criterion centers the program's participatory methods. For CLARISSA, higher-quality evidence would include system actors not only providing their perspectives, but also engaging directly in analysis and drawing conclusions about how change was emerging in the system.

- *Triangulation.* Building on common standards in qualitative research and including the need to understand systems dynamics, triangulation was considered an important way to look across the different methods to explore phenomena from various perspectives and build a robust narrative for how change was emerging and for whom.

Deliberation surfaced different perspectives on using the term "representativeness" to codify

[2] See Apgar and Snijder (2021) for an explanation of the causal hotspot practice as a way to zoom in and unpack specific causal packages to prioritize where evaluation can add most value.

[3] See Apgar et al. (2024b) for more on the findings from the evaluation.

the central principle of meaningful participation. Some colleagues felt the term would be misunderstood to suggest the use of a representative sample. As a result, greater attention was placed on contextualizing all criteria to fit the program's values on participation and complexity (CLARISSA, 2023). Two further criteria were agreed as appropriate for quality in evaluative judgements that would result from the contribution analysis design:

- *Plausibility.* The design called for careful attention to causal pathways that could explain how and why change was emerging and for whom. Plausible contribution claims depend on a clear and logical explanation of the causal steps between the participatory intervention and observed outcomes.

- *Uniqueness.* This was interpreted within the contribution analysis approach as the specificity with which a causal explanation included the effect of the CLARISSA intervention on the broader process of change. A higher-quality explanation would allow more nuanced contribution claims to be built from the evidence.

For each of the criteria, the team then discussed the levels (from 1 to 5). As with any rubric, these became qualitative descriptors of what performance on each level would look like, worded in a way that the levels would be clearly distinct. For the transparency, triangulation, plausibility, and uniqueness rubrics, we adapted the wording from previously developed rubrics by Aston (2020) to fit within the context of CLARISSA. Given there was no previously available rubric for representativeness, as a team we developed and refined what this might look like at each level and developed the descriptors for the criteria. As an example, for the representativeness rubric, the distinction between the levels was based on the extent to which the participants were involved in data collection and analysis processes and how much agency they had in the process. The difference between Level 3 and Level 4 was that participants needed to be involved in the analysis process to reach Level 4. The difference between Levels 4 and 5 was that there needed to be high levels of agency among participants throughout the whole research process, where they had ownership over certain parts of the data collection and analysis to reach Level 5. Furthermore, given participants were not a homogeneous group, we also included that the highest level (5) would be rated if the evidence contained contradictory views, as this way it would truly reflect the heterogeneous nature of the participants and the system itself, whereas in Level 4, the viewpoints would be more aggregated rather than unique.

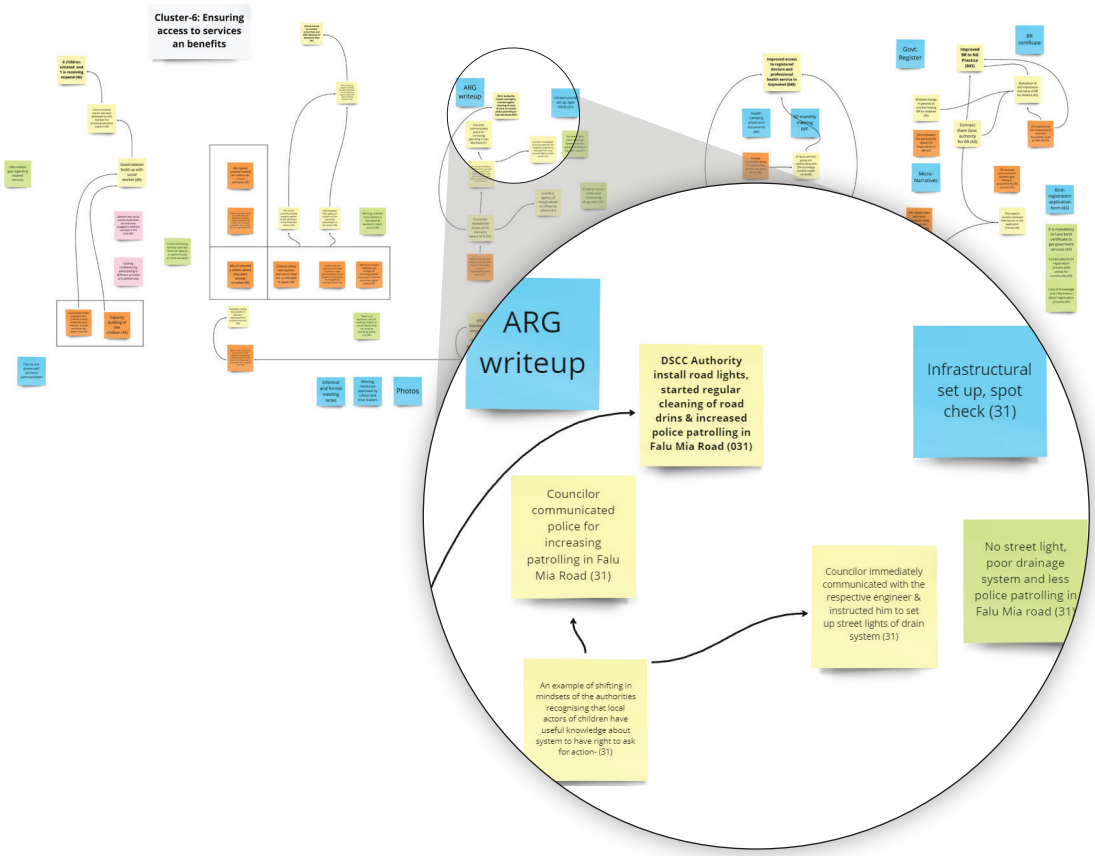
The final rubrics were published to build transparency in the way in which each performance level was contextualized and described for each of the criteria (CLARISSA, 2023).

Application of Evidence Rubrics Within Evaluation of Systems Change

The rubrics were applied in two moments of the evaluation process to assess the quality of evidence related to the causal hotspots. First, they were used within an adaptation of outcome harvesting, which was intended to document and explore how change emerged from the various systemic action research activities on the ground, including synergistic effects. The outcome evidence method used (Paz-Ybarnegaray & Douthwaite, 2017) went beyond the standard outcome harvesting practice (per Wilson-Grau, 2018) by specifically evidencing "trajectories of change" — in other words, detecting outcome patterns rather than documenting and evidencing single occurrences of outcomes in specific behaviors. As a participatory method, the program's evaluation team facilitated the generation and documentation of outcome descriptions in two rounds, in response to this question: What outcomes are emerging in system actors and domains, and what evidence do we have of how the program has contributed to them?

All collected outcomes were analyzed by the implementation team in collective analysis workshops during which outcomes were clustered by theme, location, and level of change — individual, participatory action research group, or system level — using the Water of System Change framework by Kania et al. (2018).

FIGURE 2 A Pathway Leading to Services and Benefits in Bangladesh



Collective analysis resulted in causal mapping of outcome pathways which told the contribution story and identified where program evidence backed specific causal claims. During analysis, the evidence rubrics were applied to intentionally reflect on how strong the existing evidence was in explaining the causal pathway and where gaps existed, and to design the substantiation step, during which external evaluators were commissioned to seek additional evidence and verify the program's contribution claims.

One Example

In Bangladesh, one of the pathways that led to the outcome of ensuring access to services and benefits was the result of five outcome descriptions. (See Figure 2.)

To summarize the narrative of the pathway shown in Figure 2: the local community has become well-informed about a diverse range of government and nongovernment services and benefits as a result of collaborative efforts among various system actors (reconfiguring relationships in the system), such as community groups and service providers (e.g., the partnership with the local health service provider and an advocacy initiative with the school authority). Notably, a significant shift occurred in the information flow to decision-makers in the system, driven by the active involvement of children. As a consequence of these initiatives, there has been a noticeable change in the community's mindset, fostering an increased willingness to access available services. This shift has significantly contributed to an overall improvement in the living conditions of the community. Working children now have greater educational opportunities, community members benefit from improved health care services, and the community as a whole experiences heightened

TABLE 2 Evidence Rubrics Final Assessment

Dimension	Rating	Reasoning for Rating
Transparency is about being open about where evidence for the change narrative comes from. Openness refers to who collected the data, who they were collected from and how, and how this was driven by a robust evaluation design.	5	How the data were collected and who was involved in collection and analysis is described in detail. Methodological publications discuss the development of the tools and how they were used and adapted throughout.
Triangulation relates to the use of multiple methods to build a nuanced understanding of change in complex systems; theoretical triangulation by working with multiple theories and using data from different sources and lines of evidence.	5	Evidence comes from documentation of meetings, facilitator journaling, interviews with action research group members, and their own evaluations and reflections. The implementation team was involved in making sense of the data and external evaluator substantiating, thus strengthening analyst triangulation.
Representativeness is defined based on CLARISSA's participatory ethos. It refers to the extent to which the voices of those affected by an issue are central in the evidence that is presented, and how they have participated in different parts of the process that has generated the evidence (design, data gathering, analysis, presenting).	4	Evidence is generated through participatory processes with documentation of the process. It directly includes participants making sense of their experiences through ongoing reflection sessions. Children and business owners were not involved in the final analysis of the data that informed findings in this report.
Uniqueness is about the level of confidence we have in our proposed narrative of the actual contribution of the program. It requires detailed and nuanced explanation of the link between the intervention and the outcome, identifying if there is distinctiveness of effect and by trying to rule out other factors that may have caused the outcome.	5	Evidence underpinning the causal claims made about how systemic action research generates innovative actions to tackle the worst forms of child labor is highly specific to the intervention and the outcome. It is not plausible that the actions that were generated were the result of another intervention or another process taking place at the same time as most children and business owners were not involved in other, comparable processes.
Plausibility is about the narrative of change described in the evaluation providing a clear and logical thread that follows the data.	5	Through the detailed evidence gathering in the realist evaluation, together with other methods, we have been able to develop a highly convincing account with clearly and logically signposted steps on how innovative actions were taken and influenced system dynamics.

Adapted from Apgar et al., 2024b

safety and security, thanks to robust municipal support.

As shown in Figure 2, blue stickies represent existing evidence (from the action research group writeups and a spot check on the actual infrastructure improved). Green stickies represent contextual conditions influencing the process of change.

The quality of evidence rubrics were applied to this pathway, which includes multiple system dynamics, through a facilitated process led by the evaluation team members. The purpose was for the systemic action research team to critically reflect on the quality of the existing programmatic evidence. The result illustrated that while the pathway was strong in terms of triangulation and representation, there were some weaknesses — in particular, in the plausibility of the causal explanation between the CLARISSA activities and the outcomes. This led to the development of a substantiation plan that allowed further exploration of the causal

pathway through speaking to specific system actors who shed light on the how and why of this process of change in different system dynamics.

Application of Evidence Rubrics to Final Contribution Claims

The evidence rubrics were also applied when the final contribution claims were developed along the program's multiple pathways through synthesis across the bricolage of methods. The evaluation team held sessions to deliberate and agree final scores and the reasoning for each, and the results are included in the final evaluation report (Apgar et al., 2024b). Using the original rubrics, a discussion was facilitated between team members to agree the collective reasoning for each level.

This allowed for the final assessment of all evidence presented in response to the evaluation question: How, for whom, and under what conditions did the program's systemic action research generate innovative solutions to tackle the drivers of worst forms of child labor, and what outcomes are emerging in system actors and domains? (See Table 2.)

Regarding representativeness, the team scored its performance at Level 4 and the reasoning makes explicit that participants were not involved in the final analysis, thus not fully achieving the descriptor in the original rubric of "high levels of participants' agency in the research process, analysis, and resulting actions," which would have justified a scoring of 5. In this way, the initial rubrics served as a guide for discussion and deliberation across the team, allowing critical reflection on the quality of the evidence underpinning the findings.

Conclusion and Lessons Learned

This case illustrates how more appropriate, flexible, and inclusive standards for assessing the strength of evidence in system-change efforts are achievable. Complexity-aware approaches to systems change require a greater degree of flexibility, and evaluation processes and methods need to reflect this.

The "values turn" in evaluation is an important step to re-center evaluation in what really matters for systems change. With foundations addressing ever more complex challenges, such as climate change and social and racial justice, they should more explicitly define what values should shape evaluations which help to define specifically what "quality" means in the evidence that is sought, recognizing the potential need for diversity.

The choices of methods and kinds of evidence in systems-change evaluation should be based on context specific and flexible criteria. These should be adapted to the values and questions of an evaluation. We ought not to assume that evaluators can predefine all desired outcomes. Instead, as our case study shows, assessment needs to be iterative and provide the scope to redefine boundaries as the nature of the system becomes clearer. Indeed, some criteria, such as evaluator independence or even uniqueness of contributions, may not always be appropriate, depending on what foundations are working on and the kind of changes they seek to evaluate. In the example, the choice of contribution analysis as an overarching design and the internal nature of the evaluation led to excluding independence and generalizability which are often assumed to be common standards.

While foundations need flexibility to choose what they value most, they also need some degree of structure for sensemaking. Rubrics have increasingly been seen as a useful and adaptable tool to facilitate discussion on what foundations value and how to contribute to systems change. Our case study illustrates how rubrics provide a practical architecture for a deliberative process to discuss, debate, and define what success looks like with the main evaluation stakeholders. It demonstrates the benefits of developing and applying critical appraisal tools in a participatory way with program staff centering explicitly shared values. The funder was involved early on in debating what appropriate questions and designs would be, setting up an enabling environment for the development and use of rubrics to operationalize these collective choices. In the case of

CLARISSA, given the participatory nature of the intervention itself, inclusion of community experiences was integrated through the action research processes on the ground. The specific framing of the representativeness criterion, expressing the underpinning value of inclusion, allowed the evaluation and program team to together reflect on how the participatory intention was playing out in practice. In this sense, application of the rubrics supported reflexivity of the implementation team, creating space to safely critique internal evidence and the extent to which it had been co-produced with system actors. We see this as an important step on the journey to inviting other stakeholders into an evaluation process, recognizing the complexities and power relationships that need to be navigated as we shift toward even more inclusive practice.

The case further shows that some flexibility in the rubrics used was important because it enabled the evaluation stakeholders to have robust and open conversations about quality in the face of complexity and unpredictability of causal pathways. This invites us to consider at what point in a collaborative evaluation process of complex change should the specific descriptors in rubrics become fixed, to safeguard against the risk of making the standards fit the evidence emerging allowing evaluation stakeholders to game the system. These questions are driving ongoing reflections within the communities of practice of which we are a part, enabled by foundations opening up their internal processes to actively build the field of systems-change evaluation.

References

Apgar, M., Hernandez, K., & Ton, G. (2020, September). *Contribution analysis for adaptive management*. ODI. https://media.odi.org/documents/glam_contribution_analysis_final.pdf

Apgar, M., Alamoussa, D., Báez-Silva, A. M., Bradburn, H., Deng, A. C., Cubillos Rodriguez, E., et al. (2022, April 22). *Innovating for inclusive rigour in peacebuilding evaluation*. Institute of Development Studies. https://www.ids.ac.uk/opinions/innovating-for-inclusive-rigour-in-peacebuilding-evaluation/ (accessed 17/06/23).

Apgar, M., & Snijder, M. (2021, September 15). *Finding and using causal hotspots: A practice in the making*. Institute of Development Studies. https://www.ids.ac.uk/opinions/finding-and-using-causal-hotspots-a-practice-in-the-making/

Apgar, M., Bradburn, H., Rohrbach, L., Wingender, L., Cubillos Rodriguez, E., Báez-Silva Arias, A., et al. (2024a). Rethinking rigour to embrace complexity in peacebuilding evaluation. *Evaluation, 30*(3), pp. 408–433.

Apgar, M., Snijder, M., Paul, S., Ton, G., Prieto Martin, P., Veitch, H., et al. (2024b). *Evaluating CLARISSA: Evidence, learning, and practice*. Institute of Development Studies. https://opendocs.ids.ac.uk/articles/report/Evaluating_CLARISSA_Evidence_Learning_and_Practice/26362432?file=47892121

Aston, T. (2020). *Quality of evidence rubrics*. https://www.linkedin.com/posts/tom-aston-consulting_quality-of-evidence-rubrics-activity-6736598045133164544-mfkZ/

Aston, T., Roche, C., Schaaf, M., & Cant, S. (2021). Monitoring and evaluation for thinking and working politically. *Evaluation, 28*(1), 36–57. https://doi.org/10.1177/13563890211053028

Aston, T., & Apgar, M. (2022). *The art and craft of bricolage in evaluation* (CDI Practice Paper 24). Institute of Development Studies. Available online at https://opendocs.ids.ac.uk/articles/report/The_Art_and_Craft_of_Bricolage_in_Evaluation/26433694?file=48183547

Aston, T., & Apgar, M. (2023). *Quality of Evidence Rubrics for Single Cases*. UK Evaluation Society.

Befani, B., Ramalingam, B., & Stern, E. (2015). Introduction — towards systemic approaches to evaluation and impact. *IDS Bulletin, 46*(1), 1–6. https://doi.org/10.1111/1759-5436.12116

Bicket, M., Christie, I., Gilbert, N., Hills, D., Penn, A., & Wilkinson, H. (2020). *Handling complexity in policy evaluation — Magenta Book 2020 supplementary guide*. Centre for the Evaluation of Complexity Across the Nexus. https://www.cecan.ac.uk/news/handling-

complexity-in-policy-evaluation-magenta-book-2020-supplementary-guide/

Boruch, R., & Turner, H. (2023). Randomized field experiments: Advances in practice. In M. C. Alkin and C. A. Christie (Eds.). *Evaluation roots: Theory influencing practice* (pp. 55–66). Guilford.

Brown, M. E., & Dueñas, A. N. (2020). A medical science educator's guide to selecting a research paradigm: Building a basis for better research. *Medical Science Educator, 30*(1), 545–553. https://doi.org/10.1007/s40670-019-00898-9

Brown, W., & Rosser, W. (2023). A framework for creating systems change. *The Foundation Review, 15*(4), 50–62. https://doi.org/10.9707/1944-5660.1678

Burns, D. (2007). *Systemic action research: A strategy for whole system change*. Policy Press.

Carr, M., & Morariu, J. (2023, June 26). *Fixing broken systems is hard work — and so is understanding progress*. Walton Family Foundation. https://www.waltonfamilyfoundation.org/stories/strategy-and-learning/fixing-broken-systems-is-hard-work-and-so-is-understanding-progress

Chambers, R. (2015). Inclusive rigour for complexity. *Journal of Development Effectiveness, 7*(3), 327–35. https://doi.org/10.1080/19439342.2015.1068356

Chelwa, G. (2020). Pop developmentalism in Africa, *CODESRIA Bulletin, 1*(2020), 3–5.

Chilisa, B. (2019). *Indigenous research methodologies* (2nd ed.). Sage.

Child Labour Action-Research-Innovation in South and South-Eastern Asia. (2023). *CLARISSA's quality of evidence rubrics*. Institute of Development Studies. https://doi.org/10.19088/CLARISSA.2023.003

Coné, M. A., & Dean-Coffey, J. (2024). The practice of the Equitable Evaluation Framework™. *The Foundation Review, 15*(3), 7–12. https://doi.org/10.9707/1944-5660.1663

Donaldson, S., Christie, C., & Mark, M. (2008). *What counts as credible evidence in applied research and evaluation practice?* (1st ed.). Sage.

Downes, J., & Gullickson, A. M. (2022). What does it mean for an evaluation to be "valid"? A critical synthesis of evaluation literature. *Evaluation and Program Planning, 91*(4), 1–19. https://doi.org/10.1016/j.evalprogplan.2022.102056

Gates, E., Teasdale, R., Shim, C., & Hubacz, H. (2024, June). Whose and what values? Advancing and illustrating explicit specification of evaluative criteria in education. *Studies in Educational Evaluation, 81,* 1–11. https://doi.org/10.1016/j.stueduc.2024.101335

Green, D. (2019). *What's missing in the facilities debate*. Development Policy Centre, Australian National University. https://devpolicy.org/whats-missing-in-the-facilities-debate-20190605/

Haldrup, S. (2023). What is 'good' systems change and how do we measure it? *Medium*. https://medium.com/@undp.innovation/what-is-good-systems-change-and-how-do-we-measure-it-68bec17c4d08

Kania, J., Kramer, M., & Senge, P. (2018). *The water of systems change*. FSG. https://www.fsg.org/resource/water_of_systems_change/

King, J. (2023, March 7). Developing rubrics with stakeholders. *Evaluation and Value for Investment*. https://juliankingnz.substack.com/p/developing-rubrics

Kirkhart, K. E. (2010). Eyes on the prize: Multicultural validity and evaluation theory. *American Journal of Evaluation, 31*(3), 400–413.

Loveridge, D. (2023). Evaluating systems change using rubrics. *Medium*. https://medium.com/@donna_loveridge/evaluating-systems-change-using-rubrics-6d30f0e8248d

Lowther, K., & McKegg, K. (2023, February 22). How do we define and create rigour in evaluation in complex environments? *Medium*. https://medium.com/centre-for-public-impact/how-do-we-define-and-create-rigour-in-evaluation-of-complex-environments-cd51214a0927#:~:text=By%20interacting%2C%20listening%2C%20and%20questioning,a%20result%20of%20everyone's%20efforts.

Lund, T. (2021). A revision of the Campbellian validity system. *Scandinavian Journal of Educational Research, 65*(3), 523–535. https://doi.org/10.1080/00313831.2020.1739126

Lynn, J., & Apgar, M. (2024). Exploring causal pathways amid complexity: Understanding when and how causality can be made visible. In K. E. Newcomer & S. W. Mumford (Eds.), *Research handbook on program evaluation* (pp. 304–325). Edward Elgar.

Lynn, J., & Coffman, J. (2024). Passing in the dark: Making visible philanthropy's hidden and conflicting mental models for systems change. *The Foundation Review, 16*(1), 142–160. https://doi.org/10.9707/1944-5660.1700

Lynn, J., Stachowiak, S., & Coffman, J. (2021). Lost causal: Debunking myths about causal analysis in philanthropy. *The Foundation Review, 13*(3), 17–29. https://doi.org/10.9707/1944-5660.1576

Mayne, J. (2008). *Contribution analysis: An approach to exploring cause and effect* (Institutional Learning and Change Brief No. 16). ILAC. https://cgspace.cgiar.org/handle/10568/70124

Mosley, J. M., Quarles, L. W., & Williams, J. L. (2024). Learning, unlearning, and sprinkling in: Our journey with equitable evaluation. *The Foundation Review, 15*(3), 47–59. https://doi.org/10.9707/1944-5660.1666

Nutley, S., Powell, A., & Davies, H. (2013). *What counts as good evidence?* Alliance for Useful Evidence. https://media.nesta.org.uk/documents/What-Counts-as-Good-Evidence-WEB.pdf

Paz-Ybarnegaray, R., & Douthwaite, B. (2017). Outcome evidencing: A method for enabling and evaluating program intervention in complex systems. *American Journal of Evaluation, 38*(2), 275-293.

Preskill, H., & Lynn, J. (2016, February 1). *Redefining rigor: Describing quality evaluation in complex adaptive settings*. FSG. https://www.fsg.org/blog/redefining-rigor-describing-quality-evaluation-complex-adaptive-settings/

Puttick, R., & Ludlow, J. (2013). *Standards of evidence: An approach that balances the need for evidence with innovation*. Nesta. https://media.nesta.org.uk/documents/standards_of_evidence.pdf

Raimondo, E. (2023). *The rigor of case-based causal analysis: Busting myths through demonstration* (IEG Methods and Evaluation Capacity Development Paper Series). World Bank. https://ieg.worldbankgroup.org/sites/default/files/Data/Evaluation/files/methods_paper-case_based.pdf

Schwandt, T. (2008). Toward a practical theory of evidence for evaluation. In S. I. Donaldson, C. A. Christie, & M. M. Mark (Eds.), *What counts as credible evidence in applied research and evaluation practice?* (1st ed.) (pp. 197–212). Sage.

Schwandt, T., & Gates, E. (2021). *Evaluating and valuing in social research* (1st ed.). Guilford.

Specialist Unit for Review Evidence. (2018). *Questions to assist with the critical appraisal of qualitative studies*. https://www.cardiff.ac.uk/specialist-unit-for-review-evidence/resources/critical-appraisal-checklists

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R., & Befani, B. (2012). *Broadening the range of designs and methods for impact evaluations* (Working paper No. 38). Department for International Development, United Kingdom.

White, H. (2019). The twenty-first century experimenting society: The four waves of the evidence revolution. *Palgrave Communications, 5*(1). https://www.nature.com/articles/s41599-019-0253-6

Wild, L., & Buffardi, A. (2019). *Making adaptive rigour work: Principles and practices for strengthening MEL for adaptive management*. Overseas Development Institute. https://odi.org/en/publications/making-adaptive-rigour-work-principles-and-practices-for-strengthening-mel-for-adaptive-management/

Wilson Grau, R. (2018). *Outcome harvesting: Principles, steps, and evaluation applications*. Information Age Publishing.

Thomas Aston, Ph.D., *an independent consultant, has 17 years' experience working in the international development sector. He specializes in theory-based and participatory approaches to evaluation and is a member of the Causal Pathways Initiative. Correspondence regarding this article can be addressed to him at thomasmtaston@gmail.com.*

Marina Apgar, Ph.D., *is a research fellow at the Institute of Development Studies and a member of the Causal Pathways Initiative.*

Mieke Snijder, Ph.D., *is a research fellow in the participation, inclusion, and social change cluster at the Institute of Development Studies.*

Tom Zwollo, M.Sc., *is a monitoring, evaluation, accountability, and learning specialist for TeamUp Netherlands at Save the Children Netherlands.*