Quality of Evidence Rubrics for Single Cases

Tom Aston and Marina Apgar, 2023

Introduction

At the heart of evaluation is the need for causal inference – this requires making a claim about a change. For any claim of change, it is important to consider the quality of evidence underpinning that claim. When using theory and case-based approaches to evaluation, we are exploring how an intervention has contributed to a change, directly or indirectly and often in relationship to other causal factors. In Thomas Schwandt's (2007) *Dictionary of Qualitative Inquiry*, for example, evidence is 'information that has a bearing on determining the validity of a claim.' Therefore, the focus is on the "probative value" of evidence - how much the evidence makes a particular explanation better or worse (Ribeiro, 2019).

In this document, we provide **guidance** and a set of **rubrics to assess the quality of evidence** in relation to single cases related to a particular outcome.

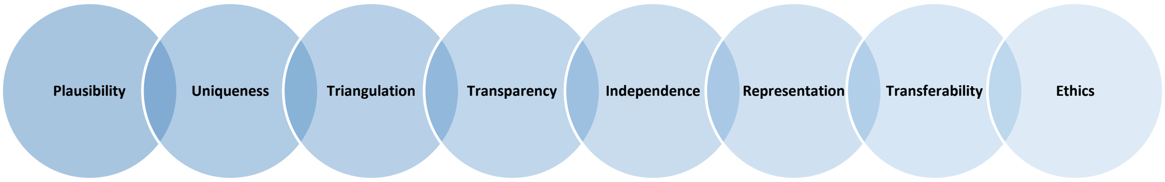
Rubrics are a form of qualitative scale that include:

- **Criteria:** the aspects of quality or performance that are of interest, e.g., timeliness.
- **Standards:** the level of performance or quality for each criterion, e.g., Poor/adequate/good.
- **Descriptors:** descriptions or examples of what each standard looks like for each criterion of the rubric (see Green, 2019; Aston, 2020a; King, 2023).

We bring together numerous evidence assessment methods and tools for causal inference (Pawson, 2007; Puttick and Ludlow, 2013; DFID, 2014; Vaca, 2016; Steadman-Bryce, 2017; Bond, 2018; CASP, 2018; SURE, 2018; Ramalingam *et al.* 2019; JBI, 2020; Gough, 2021). Because this guidance is designed for single case explanations, it focuses mainly on how to strengthen internal validity within a particular case (i.e., the extent to which a piece of evidence supports a claim about cause and effect), and where cases are likely to primarily rely on qualitative data. These evidence rubrics should, therefore, be appropriate to support various theory-based or case-based methods (such as Contribution Analysis or Realist Evaluation).

While this document does not offer full guidance for assessing external validity it does include one rubric on transferability which is more appropriate when using methods that account for context within causal claims. It also does not offer guidance on assessing the quality of an evidence base as a whole at portfolio level or across a body or research.

Below we explain 8 key criteria for assessing standards of case-based evidence: (1) **plausibility**; (2) **uniqueness**; (3) **triangulation**; (4) **transparency**; (5) **independence,** (6) **representation**, (7) **transferability**, and; (8) **ethics**. We have chosen these eight because they are common across numerous evidence assessment methods.



The rubrics proposed below all have 5 levels. Ultimately, it is up to monitoring and evaluation teams themselves to decide on what is the desired level required, but in most cases, it is reasonable to aim for level 3 to ensure credibility.

1) Plausibility

At the most basic level, the data and narrative of change should be clearly presented and the association between intervention and outcome ought to be plausible.[1] High quality studies or evaluations tend to provide a **clear, logical thread** (Puttick and Ludlow, 2013). Narratives, therefore, need to signpost the reader through the key steps and clearly explain the relationship between the intervention and the change (otherwise known as congruity). The **timing** of the outcome needs to make sense in relation to the intervention (see Beach and Pedersen, 2019). Claims of contribution and effect should be reasonable, and conclusions drawn should clearly follow the data. Below you can find a rubric to assess potential levels of plausibility:

Table 1. Plausibility Rubric

1	2	3	4	5
Unclear, illogical, or contradictory explanation connecting intervention to outcome.	Explanation indicates a possible connection between intervention and outcome.	Explanation is clear, logical, and temporally consistent, and suggests a likely association between intervention and outcome.	Convincing explanation of how evidence connects intervention and outcome. Conclusions drawn tend to follow the data.	Highly convincing account, clearly and logically signposting key steps and specific data connecting intervention to outcome. Conclusions drawn unambiguously follow the data.

This is the first layer of assessment. As a reviewer (e.g., Monitoring and Evaluation Officer), this should help you to determine whether it is worth digging further into the details of the case (likely association – level 3). In Outcome Harvesting, for example, clarifying **what** changed, **who** changed their behaviour, **when** the change took place, and **where** the change took place should help with basic plausibility. It is also helpful to make the narrative for an outcome as Specific, Measurable, Achieved, Relevant, and Timely – SMART – as possible (see Wilson-Grau and Britt, 2013; Aston, 2020b). It can also be helpful to seek out negative evidence – i.e., evidence which is contradictory (Denzin, 1970), as this can help the investigator to uncover whether there are any basic problems with the contribution claim and are an important indicator of quality in case study research more broadly (Yin, 2003).

[1] The Department for International Development's – DFID (2014) guidance refers to this as "cogency."

2) Uniqueness

We also need to understand the causal links between an intervention and a particular outcome (i.e., how good the connection is). In case studies, the **uniqueness of this connection** is a good proxy for the strength of causal links. This is also referred to as the **distinctiveness** of effect patterns and the **specificity of association** (Scriven, 2008; Norris *et al.* 2008). Relatedly, uniqueness is an important proxy for internal validity (Cook and Campbell, 1979). In qualitative terms, uniqueness indicates the degree of confidence we may have in a proposed explanation (hypothesis), based on its level of "probative value" – i.e., quality of evidence to support the hypothesis, or not (Stedman-Bryce, 2017; Ribeiro, 2019).

Demonstrating the uniqueness of the connection between intervention and outcome helps to rule out **what else** may explain the outcome (if not the intervention), otherwise known as "rival explanations" (see Dart, 2018). In this sense, assessing "uniqueness" is essentially a "smoking gun" test in Process Tracing (see Beach and Pedersen, 2019). In Contribution Analysis, "uniqueness" helps to focus on broader exploration of contribution, through looking at the whole causal package including external factors to determine what exactly the contribution of the intervention has been (Mayne, 2019). Below you can find a rubric to assess potential levels of uniqueness:

Table 2. Uniqueness Rubric

1	2	3	4	5
Evidence is found which disproves the claim, demonstrating another intervention caused the outcome.	The evidence provides a weak connection between the intervention and the outcome. Various confounding factors are possible. Other evidence indicates possible rival explanations.	The evidence provides an ambivalent connection between the intervention and the outcome. It is equally possible that the claim is valid or invalid.	The evidence is specific to the intervention. The outcome demonstrates a distinctive effect pattern. It demonstrates a probable connection between intervention and outcome. Alternative explanations are unlikely.	The evidence is highly specific to the intervention. The outcome demonstrates a very distinctive effect pattern, clearly connected to the intervention. Alternative explanations are implausible.

It is better to look at plausibility and uniqueness at every step in the process in a particular case, but what matters most is that you can establish a clear connection for key moments (or events) at which the intervention is argued to have made a difference – what you might call **"causal hotspots** (Apgar and Ton, 2021)."

3) Representation

Representation of stakeholders’ perspectives is a key part of how we should understand the probative value of evidence, especially when we are interested not just in what has been achieved, but who has benefited from the changes evidenced. While on one hand representation (or representativeness) may refer to how well and faithfully stakeholders’ views are represented quantitatively, on the other hand, and in the context of equity-oriented evaluation, representation also refers to whether the most relevant or priority groups have been engaged and whether their perspectives and experiences have shaped the understanding of a causal claim.

This goes beyond seeking out **multiple perspectives** from different stakeholders to check the credibility of a particular narrative or causal claim. To some degree, we want to know whether there has been adequate coverage of a particular population. This refers to whether an acceptable proportion of a specific population has been reached and consulted through interviews, surveys, or other means of data collection. Yet, the benefits of representative samples have been seriously questioned in replication studies (Coppock *et al.* 2018).

Sampling and decisions about who to involve and to what extent ought to be driven by the questions we want to answer and the ideas (or theories) about the social world we seek to investigate, anchored to a particular context. Therefore, different groups may be more or less relevant to help you to answer those questions and develop, refine, or test theory (Maxwell, 2012; Emmel, 2013). In this light, the number of units included is less important than how the insights into events and experiences are used for interpretation (and in some cases this requires engagement of stakeholders in the interpretation itself) explanation, and as support (or not) for the claims we make (Emmel, 2013). As Pawson *et al. (*2004: 20) argue, we stop looking when ‘sufficient evidence has been assembled to satisfy the theoretical need or answer the question.’

Secondly, sampling in qualitative research and evaluation is more commonly driven by saturation point (i.e., when no new information is discovered that adds to our understanding) rather than statistical representativeness. Estimates vary across studies. However, saturation point is often reached somewhere around as few as 9-17 interviews (Guest *et al.* 2006; Hennink and Kaiser, 2022) and 4-8 focus group discussions for each population group (Guest *et al*. 2017; Hennink and Kaiser, 2022). So, theoretical representation can be reached with relatively low numbers.

However, more important than the brute number of people consulted is **whether and how the perspectives and experiences of priority groups are represented and shape the findings**. Priority groups may be community members, service users, public servants, politicians, or whichever other group’s experiences you are evaluating or studying.

Table 3. Representation Rubric

1	2	3	4	5
The perspectives and/or experiences of priority groups are not included as sources of evidence.	The perspectives and/or experiences of some priority groups have been included, but those groups have not been	The perspectives and/or experiences of priority groups have been elicited indirectly through data collection by	Priority groups generate their own evidence with their own perspectives and experiences. These are aggregated and homogenised	Multiple sources of evidence generated directly by priority groups through their own data collection and analysis

	involved or consulted.	the researchers or evaluators and from their observations.	by researchers and/or evaluators, so therefore may lack nuance.	processes. Perspectives may be unique to different groups and thus reflect a variety of viewpoints.

4) Triangulation

Triangulation is a key area for evidence quality and helps ensure a degree of consistency and bias control. According to Better Evaluation (2022b), triangulation tests the consistency of findings obtained through different instruments and increases the chance to control, or at least assess some of the threats or multiple causes influencing our results.

Norman Denzin (1970, 2006) refers to **four main types of triangulation**: (1) data triangulation; (2) investigator triangulation; (3) theory triangulation, and; (4) methodological triangulation. To some degree, these types can be disaggregated further. William Shadish (1993) also emphasises: (5) coder triangulation and (6) analyst triangulation,[2] and we may also reasonably add (7) perspectival triangulation, which may overlap with data triangulation, but not necessarily (see representation below). It may not be feasible or worthwhile to achieve all of these in a single study, but they offer a helpful reminder of relevant dimensions.

Theoretical triangulation can partly be covered by rival hypothesis testing which we can achieve through the uniqueness rubric above, because a rival hypothesis may involve an alternative theoretical model or scheme to interpret the phenomenon studied. Every data gathering type is potentially biased (Denzin, 1970), so **methodological triangulation** (combining different methods) or bricolage (combining parts of different methods) can be helpful to diversify lines of enquiry and data collection (Aston and Apgar, 2022). **Investigator triangulation** (including multiple researchers/evaluators/observers as part of an investigation) can help to increase the potential reliability of findings and enhance the opportunity for different interpretations of the data collected (Denzin, 1970). This can include involvement of stakeholders themselves in interpretation (as is covered by the representation rubric). **Coder and analyst triangulation** (who classifies and who assesses the data) can also potentially help make analyses more reliable (Shadish, 1993). And **perspectival triangulation** can help to ensure that the experiences and perspectives of priority groups are adequately and accurately reflected in explanations.

Data triangulation can involve triangulation of data from different time periods, locations, and people. However, this more commonly refers to **multiple sources** (e.g., multiple interviewees, multiple questionnaires) or **multiple lines of evidence** (i.e., different source types interviews *and* questionnaires) (see Denzin, 1970). Data sources can be primary and secondary, so they may be existing data sources or those that are created through the investigation. You may seek corroboration between administrative, testimonial, and observational sources, and data may be obtained through **different methods** (interviews, observations, questionnaires, etc.).

Not all sources are necessarily of equal value. Some voices may be more important than others due to their proximity to or perspective on the change reported (e.g., eyewitness accounts vs. hearsay). Given that all sources of evidence have some degree of bias (see White and Philips, 2012), it is important to consider which can partially, or fully, corroborate your proposed narrative of change. For this reason, it is often important to seek out **multiple perspectives** from different stakeholders to check the credibility of a particular narrative or explanation. Where feasible, it can also be helpful to triangulate across **different studies** and tools to check for consistency of findings. These may be within an evaluation (e.g., baseline, mid-line, end-line assessment) or from other forms of research.

Investigators, coders, and analysts may or may not be the same people. An investigator may refer to a data collector, coder, and analyst. It is also possible to have a separate analyst and synthesist, where the synthesist writes up the findings of the analyst (Pasenen *et al*.

[2] For Denzin (1970: 303) both coder and analyst are included under investigator triangulation.

2018). There are benefits to having the same and different people play these roles. Having the same people can potentially enhance consistency and accuracy. Yet, having different people may enhance potential reliability and independence.

Particularly when evaluations are participatory and do not include sufficient reflexivity, proximity and confirmation **biases** may be stronger. Participants might develop supposed causal explanations which reflect the greatest intensity of their efforts (i.e., intervention) rather than where there is necessarily evidence of a connection to the proposed change if they are not carefully facilitated to use critical reflection throughout (Wadeson *et al.* 2020). This is particularly the case for testimonial evidence. Biases are unavoidable, but triangulation is one way to limit potential respondent biases.[3] On the following page you can find a rubric to help assess levels of triangulation:

Table 4. Triangulation Rubric

1	2	3	4	5
No evidence corroborates the connection between intervention and outcome. Other evidence contradicts the proposed connection.	A single source of evidence supports the claim (i.e., the connection between intervention and outcome).	Multiple lines of evidence (i.e., source types)[4] corroborate the connection between intervention and outcome.	Multiple lines of high-quality evidence corroborate the connection between intervention and outcome.[5]	Multiple lines of evidence across different studies corroborate the connection between intervention and outcome.

[3] Testimonial evidence is typically very important for qualitative studies, especially if there are gaps in documentary or open-source evidence. However, we may not necessarily be looking to achieve representativeness, but rather to shed light on different perspectives and improve our understanding of the change we seek to explain. Simply increasing the number of interviews has diminishing returns in terms of revealing *new* qualitative information. While contested as potentially misleading (see Braun and Clarke, 2019), it is commonly estimated that saturation point of around 90% is typically found somewhere between 6 and 16 interviews for each stakeholder group (Guest *et al.* 2006, 2020). A recent systematic review finds that saturation point is reached between 9 and 17 interviews (Hennink and Kaiser, 2022). In this sense, seeking more perspectives can sometimes be helpful, but not always.

[4] E.g., interviews, focus groups, surveys, observation,

[5] High quality evidence refers to evidence with high "probative value." See section on "uniqueness."

5) Transparency

Transparency and openness underpin strong causal claims and all evaluation and research processes. This entails that we know as much as possible about where the evidence comes from, who collected it, and how it was collected. For this, some details should be provided on what the sources of data are, the methods used, results achieved, and any key limitations in the data or conclusions (CASP, 2018; SURE, 2018; JBI, 2020). High quality accounts should also be self-critical, **identifying limitations**, exploring **alternative interpretations** of the analysis and potential **rival explanations** linked to other factors (i.e., uniqueness). This makes transparency all the more important. Below you can find a rubric to assess potential levels of transparency:

Table 5. Transparency Rubric

1	2	3	4	5
It is unclear what evidence supports the claim.	Evidence has been identified, but not clearly explained.	Various sources of evidence are clearly identified and explained.	Sources of evidence and data collection methods are clearly explained. Data limitations and alternative interpretations are clearly discussed.	Sources of evidence and data collection methods are clearly explained. Data limitations and alternative interpretations and the plausibility of alternative explanations are clearly discussed. Data collection protocols available.

Finally, we can look at how raising the degree of independence can potentially increase the quality of evidence underpinning causal claims.

6) Independence

Independence is a key area for standards of evidence especially in the case of impact evaluations which tend to be implemented by expert evaluators that are commissioned because they are independent. Yet, even if an evaluator is not connected to the intervention, there are also biases they hold that may need to be navigated. Independence is also important in the case of observational methods which have various potential sources of bias (CASP, 2018; SURE, 2018; JBI, 2020). Most often, this refers to testimonial evidence where respondents have a personal connection to the intervention, but it also refers to project records, archival evidence, and even public speeches.

Self-reported data are usually considered lower quality due to several potential biases. Staff tend to want to keep their jobs. They will tend to focus on describing events to which they and their organisation were connected (proximity bias) and which make themselves look good (self-serving bias), whether these were causally significant or not. Indeed, their explanations often tend to diminish the role and contribution of others in the process.

Data from partners also tends to have a number of limitations in terms of independence. Partners have similar incentives linked to contract renewal. They will thus tend to say things which they believe are acceptable to the organisation in question (social acceptability bias), whether such answers are truthful or not. For similar reasons, they will also tend to provide accounts which confirm what the organisation believes to be true or wants to be true (confirmation bias).

Given the limitations mentioned above, **stakeholders formally outside of the initiative are often consulted to corroborate** the connection between intervention and outcome. **Evaluators also have a number of potential biases** which limit their independence. Like partners, they may also have contract renewal incentives. They may also be overly focused on the intervention, missing out key information about other contextual factors that influence change (intervention bias). They may even be the friends or family of those who implement the intervention (friendship bias). **Communities, local government,** and **private sector actors are typically less commonly connected** to those who implement the intervention, although this depends on how participatory the intervention is in practice. However, if they are known to the intervention, they too may display courtesy bias (due to politeness), self-serving bias (they wish to continue to receive funds in their location), or social acceptability bias (due to social norms).

Generally speaking, actors who have **first-hand experience of events** tend to have the most relevant perspective on those events. Indeed, in equity-oriented evaluation, the lived experience of the marginalised communities the intervention (and evaluation) serve are important sources of evidence. However, where possible, it is good to ensure that these actors do not have clear connections to the intervention for all of the reasons mentioned above. For each of these actors, it is important to signpost who they are and what their potential connections may be. For example, it is worth specifying the source in the definition of evidence – e.g. “An article in a left-leaning newspaper reported X (Fairfield and Charman, 2017), or staff member reported Y.” Likewise, it is worth noting that public statements typically contain more positive bias. Therefore, confidential sources are generally preferable as these are generally less subject to social acceptability and self-serving biases (Beach and Pedersen, 2019). Below you can find a rubric to assess potential levels of independence:

Table 6. Independence Rubric

1	2	3	4	5
Evidence is self-reported. Sources are from the project/programme and known to have significant biases and strong incentives to potentially misrepresent the events.	Evidence is self-reported. Primary and/or secondary data indicate a potential lack of independence and number of potential biases.	Evidence may be collected by third parties, partners or collected by independent evaluators. Issues of potential bias are unknown.	Evidence is collected by independent evaluators without clear connections to the intervention. Sources of potential bias are clearly signposted, and efforts have been made to limit these.	Evidence is collected by independent evaluators without clear connections to the intervention. Sources of potential bias are clearly signposted and considerable efforts have been made to limit these.

7) Transferability

External validity historically referred to the **approximate validity of an assumed causal relationship can be generalised to different types of persons, settings, and times** (Cook and Campbell, 1979). There have been recent critiques of generalizability based on sample sizes or exclusively based on population (Coppock *et al.* 2018; Burchett *et al.* 2020). Some have suggested that generalizability, assessments should focus on understanding an intervention's mechanism of a why or how an intervention was effective; what is required is a justification and a warrant for any claim of even modest generalization. Yet, these are often unconvincing (AREA, 2006; Burchett *et al.* 2020). Indeed, claims vary in their breadth and depth (Gough, 2021). So, it is not always clear what the scope of generalizations really are. Understanding the mechanisms of action and modifiers – the way in which an intervention interacts with its context to lead to an effect – is increasingly seen as important (Bates and Glennerster, 2017; Burchett *et al.* 2020: 2). This helps us to understand how and why an intervention may have been effective in context.

For decades, therefore, social researchers and evaluators have pointed out that **context matters for whether an intervention/project/programme will work or not in different locations, time periods, and for different types of persons** – i.e., populations (Cronbach, 1975; Lincoln, and Guba, 1985). As Cronbach (1975: 124 – 125) argued, 'when we give proper weight to local conditions, any generalization is a working hypothesis, not a conclusion.' Hence, in practice, if any, interventions/projects/programmes will work in exactly the same way everywhere, so it is highly unlikely that they will achieve exactly the same effects for different types of persons, in different types of settings, and at different moments in time. In other words, there are rarely if ever absolutes (Lincoln, and Guba, 1985).

Increasingly, **evaluators have asked not merely what works and to what extent, but how and why it worked, for whom it worked and in what contexts it worked** (Pawson and Tilly, 1997; Punton *et al.* 2020). This gave rise to the concept of transferability as a more accurate context-sensitive alternative to generalization (Aston *et al.* 2021). The concept has been an established principle for assessing rigour in qualitative research for many years. As Lincoln and Guba (1985) discussed, we can tell whether a working hypothesis in context A might be applicable to context B based on the degree of transferability, or level of fit, between context A and context B. This perspective emphasised the level of congruence between the two contexts and assumed that uncontrolled factors matter in whether effects will take place or not (see also Lincoln and Guba, 1989). The same reasonably applies to congruence between different types of persons being compared in context A and context B. Where persons (or populations) are significantly different, it is less likely that they will perceive and respond to an intervention/project/programme in the same way. Assessing similarity in people can itself be problematic and reinforce unhelpful power dynamics, particularly when taking an intersectional approach in programming that aims to reach excluded populations. Thus, understanding if there is congruence across types requires careful analysis using an intersectional approach.

Recent work by the Centre of Excellence for Development Impact and Learning (CEDIL) draws attention to the **importance of potential moderating factors** or what Nancy Cartwright refers to as "support factors" – these are the factors which are either part of intervention design or wider context that support the assumed causal process to operate (Davey *et al.* 2018; Masset and White, 2019; Cartwright *et al.* 2020, and Cartwright, 2020; White, 2022; White, forthcoming)." There are also evidently barriers to implementation. Cartwright (2020) calls these "derailers," and these make the implementation of an intervention less feasible, and the likelihood of achieving intended outcomes less likely. More commonly, these are referred to as intervention or contextual assumptions (see Williams, 2017). They are typically outlined in a theory of change. In Contribution Analysis these are part of causal link assumptions that are interrogated through evaluation (Mayne, 2019).

Davey *et al.* (2018) point out that **adaptation is often needed so that interventions are feasible and applicable to new contexts**. Given this, they suggest that in transferring an intervention we should wish to preserve fidelity of function (the way in which the intervention is intended to generate outcomes), not fidelity of form (activities, materials, delivery). They further note that implementation feasibility differs between different settings depending on the structures and resources in place that support or act as barriers to delivery. Therefore, acknowledging transferability as a quality criterion can draw attention to including more useful explanations in support of such adaptations.

In sum, therefore we underscore **four main areas to assess the level of potential transferability** between intervention/project/programme's delivered between one context and another. This entails assessing the degree of similarity (or "fit") between:

- Populations (understood through an intersectional analysis) compared in contexts A and B and their capacities, opportunities, and motivations;
- Intervention/project/programme features proposed and their appropriate fit in contexts A and B (function, not form);
- Feasibility of implementation in contexts A and B and time periods 0 and 1 (i.e., barriers to implementation);
- Salient contextual support factors to the intervention/project/programme in contexts A and B, and time periods 0 and 1.

Below you can find a rubric to assess potential levels of transferability:

Table 7. Transferability Rubric

1	2	3	4	5
Key support factors are unknown. No similarities between populations, interventions, and contexts compared at the time.	Key support factors are known, but not clearly explained. Few similarities between populations, interventions, and contexts compared at the time.	Key support factors have been clearly identified and explained. There are limited similarities between populations, intervention characteristics and contexts compared.	Key support factors have been clearly identified and explained. There are some similarities between populations, intervention characteristics, and contexts compared.	Key support factors have been clearly identified and explained. There are significant similarities between populations, intervention characteristics, and contexts compared.

These features are not static, but rather dynamic and may evolve over time, so it is also important to examine the process of implementation and how the intervention/project/programme interacts with context.

8) Ethics

At the core of any evidence assessment should be how that evidence is collected, processed, analysed, and reported back to those whose time and knowledge was used within any study. In research, it is increasingly common for researchers to be required to engage with ethical considerations at design and throughout the research process. This is often through use of an **institutional review board** (IRB) or independent ethics committee (IEC), ethical review board (ERB), or research ethics board (REB) which are tasked with assessing levels of ethical risk and ensuring mitigation measures are in place. This is less common in evaluation. However, many organizations have procedures for safeguarding of vulnerable populations, and there are also legal standards around data protection such as the General Data Protection Regulation (GDPR).

Various organisations have developed **ethical guidelines** to cover the conduct of evaluation, including the responsibilities of those conducting and managing evaluations.[6] The United Nations Evaluation Group (UNEG) (2020) have specific ethical guidelines. They emphasise integrity and accountability, and respect, and beneficence as underlying ethical principles and offers a checklist. These principles also refer to several criteria mentioned in this guidance such as independence, transparency, and fair representation. Here we are talking of potential conflicts of interest and sharing information particularly with populations affected by the evaluation. This also requires attention to questions that may arise, particularly in relation to any potential misconduct.

Research quality bodies such as the Specialist Unit for Review Evidence (SURE) and the Critical Appraisal Skills Programme (CASP) and the Joanna Briggs Institute (JBI) also provide some guidance on case-based or qualitative evidence which offer some checklist questions relevant to ethics (CASP, 2018; SURE, 2018; JBI, 2020). These checklists emphasize issues such as confidentiality, conflicts of interest, and whether ethical issues are discussed or not, and ethical approval processes.

Conflicts of interest may arise when researchers or evaluators have connections to the project. These should be clearly identified and addressed, as they will affect how researchers/evaluators design their study, collect data, and write up their findings.

Ethical approval may or may not be through a formal body, but a formal process by an independent body can add value through an additional layer of scrutiny on things that may not be directly considered by the study team. It can therefore be helpful to define specific protocols for how data will be collected and used. However, these are often insufficient on their own. In evaluations that require deeper engagement with participants, "situated ethics" (or ethics in practice) offer additional guidance around paying attention to place and context and power relationships between the evaluator and participants.[7]

In additional to any formal or informal ethical approval process, researchers and evaluators should always **gain the consent of study populations**, either verbally or in writing.

Confidentiality is a critical component of research and evaluation ethics. Stakeholders consulted ought to know how their information will be used and with whom, and this knowledge will often affect the answers they provide and the accuracy and completeness of

[6] Better Evaluation offer a set of resources from the American Evaluation Association (AEA), Australasian Evaluation Society (AES), and the African Evaluation Association (AfrEA) (Better Evaluation, 2022a).

[7] Situated ethics is most commonly applied in qualitative research traditions that require engagement in field contexts (see e.g., Guillemin, M., and Gillam, L. (2004). Ethics, reflexivity, and "ethically important moments" in research. *Qualitative inquiry*, *10*(2), 261-280. Hammersley, M. (2010). Creeping ethical regulation and the strangling of research. *Sociological Research Online*, *15*(4), 123-125.)

that information. They may reveal different information if they can/cannot be identified or if they have concerns regarding how their information will be used. It is also important to communicate any potential challenges of identifiability to those consulted, and any related risks.

How information is collected affects how trustworthy it is. When researchers or evaluators collect data, it is important that they understand the local cultural context such as local customs and gender norms and be sensitive to these. Insensitivity and inappropriateness may not only cause potential harm to study populations but can also affect the accuracy and completeness of their responses. Situated ethics, and frameworks around ethics of care are useful approaches that support the practice rather than simply checking boxes in an initial assessment.

When information is solicited through unethical means, it can reduce trustworthiness. For example, if information is obtained through coercion, this will often make it less trustworthy because people may reveal untruthful or inaccurate information under coercion. Furthermore, vulnerable groups may have lower levels of literacy and language competency, so they may be less aware of issues of consent and confidentiality when they participate in interviews, focus groups, surveys, or other data collection processes.

It may be important to explicitly **discuss ethical issues** that either arise from the researchers/evaluators or study populations. Such discussion can also help the reader to put description and analysis in the study in context. However, researchers and evaluators also need to consider the benefits and harms of publishing potentially sensitive information in their studies.

It is also important to allocate sufficient resources to enable appropriate representation and treatment of stakeholder groups. This may also include adequate resources for participatory and empowerment approaches (UNEG, 2020).[8]

Table 8. Ethics Rubric

1	2	3	4	5
Ethics have not been taken into consideration and the evaluation practice has the potential to do harm by putting stakeholders at risk.	Ethics have only been indirectly considered. There may be some potential conflicts of interest, questions over fair treatment of stakeholders and/or confidentiality.	Populations affected by the evaluation have been directly consulted, formal consent has been provided and confidentiality secured.	Ethical approval has been sought from review boards and communities. Specific procedures are in place and ethical issues explicitly discussed.	Ethical approval has been sought from review boards and communities, and evaluators have considered situated ethics in their practice.

It should be noted that not all the rubrics listed above are not necessarily of equal value. For example, uniqueness can add more weight than plausibility when trying to increase confidence in contribution claims (see Stedman-Bryce, 2017). It can therefore be important to weight these five dimensions differently (see Davies, 2020 on weighted checklists). However, the relative importance of each of the aforementioned components is a matter of

[8] There may be various other ethical considerations which may not affect the quality of evidence directly but are worth considering seriously such as secure and safe data collection, storage, and use, and whether complaints mechanisms are in place, for example. Further information can be found in UNEG, 2020.

judgement and depends on what each organisation or project values. We therefore leave aggregate scoring to your judgement. Moreover, seven criteria may very well be beyond the capacity of all organisations or projects. So, you may instead want to select from these criteria those that matter most for the task at hand and potentially use aspects of the others (as relevant) elsewhere in your study. This selection process is best done at the outset and through a participatory process that considers the preferences and values of all evaluation stakeholders.

References

American Educational Research Association (AERA) Standards for Reporting on Empirical Social Science Research in AERA Publications American Educational Research Association, Educational Researcher, Vol. 35, No. 6, pp. 33–40.

Apgar, M. and Ton, G. (2021). Learning through and about Contribution Analysis for Impact Evaluation, available at: https://www.ids.ac.uk/opinions/learning-through-and-about-contribution-analysis-for-impact-evaluation/

Aston, T. (2020a). “Rubrics as a Harness for Complexity,” available at: https://www.linkedin.com/pulse/rubrics-harness-complexity-thomas-aston/

________(2020b) “How to be Smarter about Narrating Behaviour Change,” available at: https://www.linkedin.com/post/edit/6617198441892319232/

Aston, T., Roche, C., Schaaf, M., and Cant, S. (2022). Monitoring and evaluation for thinking and working politically. *Evaluation*, *28*(1), 36–57.

Aston, T. and Apgar, M. (2022). The Art and Craft of Bricolage in Evaluation, CDI Practice paper 24, Brighton: Institute of Development Studies, available at: https://opendocs.ids.ac.uk/opendocs/handle/20.500.12413/17709

Bates, M. and Glennerster, R. (2017). The Generalizability Puzzle, Stanford Social Innovation Review.

Beach, D. and Pedersen, R.B. (2019). *Process-Tracing Methods: Foundations and Guidelines*, Second Edition. Ann Arbor MI: University of Michigan Press.

Better Evaluation (2022a). Ethical Guidelines, available at: https://www.betterevaluation.org/methods-approaches/methods/ethical-guidelines

Better Evaluation (2022b). Triangulation, available at: https://www.betterevaluation.org/methods-approaches/methods/triangulation

Bond (2018). “Principles and checklist for assessing the quality of evidence,” available at: https://www.bond.org.uk/resources/evidence-principles

Braun, V. and Clarke, V. (2019). “To saturate or not to saturate? Questioning data saturation as a useful concept for thematic analysis and sample-size rationales,” *Qualitative Research in Sport, Exercise and Health*.

Burchett, H. E. D., Kneale, D., Blanchard, L., and Thomas, J. (2020). When assessing generalisability, focusing on differences in population or setting alone is insufficient. *Trials*, 21(1), Article 286.

Cartwright, N. (2020). Using middle-level theory to improve programme and evaluation design. CEDIL Methods Brief. Oxford: CEDIL.

Cartwright, N., Charlton, L., Juden, M., Munslow, T. and Williams, R. B. (2020). Making predictions of programme success more reliable. CEDIL Methods Working Paper. Oxford: Centre of Excellence for Development Impact and Learning (CEDIL).

Cook, T. and Campbell, D. (1979). *Quasi experimentation: Design and Analysis Issues for Field Settings*, Chicago: Rand McNally.

Coppock, A. Leeper, T. and Mullinix, K. (2018). Generalizability of heterogeneous treatment effect estimates across samples, PNAS, Vol. 114, No. 49.

Critical Appraisal Skills Programme – CASP (2018). CASP Checklist: 10 Questions to Help you make sense of a Qualitative Research.

Cronbach, L. (1975). Beyond the Two Disciplines of Scientific Psychology, American Psychologist, 30, pp. 116 – 127.

Dart, J. (2018). “The What Else Tool: A Basic way to Strengthen your Impact Claims and Avoid having Egg on your Face!,” available at: https://www.clearhorizon.com.au/all-blogposts/the-what-else-tool-a-basic-way-to-strengthen-your-impact-claims-and-avoidhaving-egg-on-your-face.aspx

Davey C, Hargreaves J, Hassan S, Cartwright N, Humphreys M, Masset E, Prost A, Gough D, Oliver S, Bonell C. (2018). Designing Evaluations to Provide Evidence to Inform Action in New Settings, CEDIL Inception Paper No 2: London.

Davies, R. (2020). “Rubrics? Yes, but,” available at: https://mandenews.blogspot.com/2020/04/rubrics-yes-but.html

Denzin, N. (1970). The Research Act in Sociology, Butterworth & Co: London.

Denzin, N. (2006). *Sociological Methods: A Sourcebook.* Aldine Transaction.

Department for International Development – DFID (2014). “Assessing the Strength of Evidence: How to Note,” available at: https://www.gov.uk/government/publications/how-to-note-assessing-the-strength-of-evidence

Emmel, N. (2013). Sampling and Choosing Cases in Qualitative Research: A Realist Approach, SAGE Publications Ltd.

Fairfield, T. and Charman, A. (2017). “Explicit Bayesian Analysis for Process Tracing: Guidelines, Opportunities, and Caveats,” *Political Analysis*, Vol. 25, No. 3, pp. 363-380.

Gough, D. (2021). Appraising Evidence Claims, Review of Research in Education March 2021, Vol. 45, pp. 1– 26.

Green, D. (2019). What’s Missing in the Facilities Debate? DevPolicy, available at: https://devpolicy.org/whats-missing-in-the-facilities-debate-20190605/

Guest, G, Bunce, A. and Johnson, L. (2006). “How Many Interviews Are Enough? An Experiment with Data Saturation and Variability,” *Field Methods*, Vol. 18, No. 1, pp. 59- 82.

Guest, G., Namey, E., and McKenna, K. (2017). How Many Focus Groups Are Enough? Building an Evidence Base for Nonprobability Sample Sizes. *Field Methods*, *29*(1), 3–22.

Guest G, Namey E, Chen M (2020). A Simple Method to Assess and Report Thematic Saturation in Qualitative Research, PLoS ONE 15(5):e0232076.

Guillemin, M. and Gillam, L. (2004). Ethics, reflexivity, and “ethically important moments” in research. *Qualitative inquiry*, *10*(2), 261-280.

Hammersley, M. (2010). Creeping ethical regulation and the strangling of research. *Sociological Research Online*, *15*(4), 123-125.

Hennink, M and Kaiser, B. (2022). Sample sizes for saturation in qualitative research: A systematic review of empirical tests, *Social Science & Medicine*, Volume 292.

Joanna Briggs Institute – JBI (2020). Checklist for Qualitative Research, Joanna Briggs Institute.

King, J. (2023). Different Kinds of Rubrics, Medium, available at: https://juliankingnz.substack.com/p/different-kinds-of-rubrics

Lincoln, Y. and Guba, E. (1985). *Naturalistic Inquiry*, Sage Publications Inc: Beverly Hills, California.

Lincoln, Y. and Guba, E. (1989). *Fourth Generation Evaluation*, Sage Publications Inc: Beverly Hills, California.

Mayne, J. (2019). Revisiting contribution analysis. *Canadian Journal of Program Evaluation*, *34*(2).

Masset E. and White H. (2019). To Boldly Go Where No Evaluation Has Gone Before: The CEDIL Evaluation Agenda. CEDIL Paper: London.

Maxwell, J. (2012). A Realist Approach for Qualitative Research, Sage, Inc.

Norris, R. Nichols, S. Ransom, G. Liston, P. Barlow, A. Mugodo, J. (2008). Causal Criteria Methods Manual: Methods for Applying the Multiple Lines and Levels of Evidence (MLLE) Approach for Addressing Questions of Causality.

Pasanen, T. Raetz, S. Young, J. and Dart, J. (2018). Partner-led Evaluation for Policy Research Programmes: A Thought Piece for the KNOWFOR Programme Evaluation, available at: https://cdn.odi.org/media/documents/11969.pdf

Pawson, R. (2007). Realist Synthesis: Supplementary Reading 6: Digging for Nuggets: How 'Bad' Research can yield 'Good' Evidence.

Punton, M. Vogel, I. and Lloyd, R. (2016). Reflections from a Realist Evaluation in Progress: Scaling Ladders and Stitching Theory, CDI Practice Paper 18, Brighton: IDS.

Puttick, R. and Ludlow, J. (2013). Standards of Evidence: An Approach that Balances the Need for Evidence with Innovation, available at: https://media.nesta.org.uk/documents/standards_of_evidence.pdf

Ramalingam, B. Wild, L. and Buffardi, A. (2019). Making Adaptive Rigour Work: Principles and Practices for Strengthening Monitoring, Evaluation and Learning for Adaptive Management, available at: https://www.odi.org/sites/odi.org.uk/files/resource-documents/12653.pdf

Ribeiro, G. (2019). Relevance, probative value, and explanatory considerations. *The International Journal of Evidence & Proof*, *23*(1–2), 107–113.

Scriven, M. (2008). A Summative Evaluation of RCT Methodology: & An Alternative Approach to Causal Research, *Journal of MultiDisciplinary Evaluation*, Vol. 5, No. 9.

Shadish, W. R. (1993). Critical multiplism: A research strategy and its attendant tactics. *New Directions for Program Evaluation*, 1993, 13–57.

Schwandt, T. (2007). The SAGE Dictionary of Qualitative Inquiry. SAGE Publications Inc.

Specialist Unit for Review Evidence – SURE (2018). Questions to Assist with the Critical Appraisal of Qualitative Studies.

Stedman-Bryce, G. (2017). "Avoiding the Data Trap Blog 1: Mining for Data Gold," available at: https://insights.careinternational.org.uk/development-blog/avoiding-the-data-trapblog-1-mining-for-data-gold

United Nations Evaluation Group – UNEG (2020). Ethical Guidelines for Evaluation, available at: http://www.unevaluation.org/document/detail/2866

Vaca, S. (2016). 8 Steps to Inferring Causality, available at: http://www.saravaca.com/project/8-steps-causality/

Wadeson, A. Monzani, B. and Aston, T. (2020). "Process Tracing as a Practical Evaluation Method: Comparative Learning from Six Evaluations," available at: https://mande.co.uk/2020/media-3/unpublished-paper/process-tracing-as-a-practical-evaluation-method-comparative-learning-from-six-evaluations/

Williams, M. (2017). External Validity and Policy Adaptation: A Five-step Guide to Mechanism Mapping, Policy Memo, Blavatnik School of Government, University of Oxford.

Wilson-Grau R. and Britt, H. (2013). "Outcome Harvesting," Ford Foundation, MENA Office, available at: https://www.outcomemapping.ca/download/wilsongrau_en_Outome%20Harvesting%20Brief_revised%20Nov%202013.pdf

White, Howard, and Daniel Philips. (2012). Addressing Attribution of Cause and Effect in Small N Impact Evaluations: Towards an Integrated Framework, Technical Report 15 International Initiative for Impact Evaluation Working Papers.

White, H. (2022). Transferring evidence between contexts: Highlights from CEDIL annual conference.

White, H. (forthcoming). The use of middle-level theory in CEDIL-funded research studies.

Yin, R. K. (2003). Designing case studies. *Qualitative research methods*, *5*(14), 359-386.