Taking Stock of Qualitative Methods of Evaluation: A Study of Practices and Quality Criteria

Thilo Bodenstein, Achim Kemmerling

SciencesPo
LABORATOIRE INTERDISCIPLINAIRE
D'ÉVALUATION DES POLITIQUES PUBLIQUES

LIEPP Working Paper
November 2024, nº171

Taking Stock of Qualitative Methods of Evaluation: A Study of Practices and Quality Criteria

Thilo BODENSTEIN
Central European University
BodensteinT@ceu.edu

Achim KEMMERLING
Willy Brandt School of Public Policy
achim.kemmerling@uni-erfurt.de

www.sciencespo.fr/liepp/en/

How to cite this publication:
BODENSTEIN, Thilo, KEMMERLING, Achim, **Taking Stock of Qualitative Methods of Evaluation: A Study of Practices and Quality Criteria**, *Sciences Po LIEPP Working Paper* n°171, 2024-11-30.

Taking Stock of Qualitative Methods of Evaluation: A Study of Practices and Quality Criteria

Abstract

Research on evaluation has mapped the landscape of quantitative evaluation methods. There are far fewer overviews of the practice of qualitative evaluation methods. We present a meta-study of scholarly articles from five widely read evaluation research journals, examining the types of methods used and the transparency of their quality criteria. Our sample includes 50 out of about 1070 articles. First, we document a remarkable variety of qualitative methods, but some stand out: Case studies and stakeholder analysis, often combined with interview techniques. Articles rarely define and conceptualise their methods explicitly. This is understandable from a practical point of view, but it can make it difficult to critically interrogate findings and build knowledge. Finally, we find that the transparency of qualitative criteria required in the literature is not always sufficient, which can hinder the synthesis of results.

Keywords: Research on Evaluation, Meta-Study, Qualitative Methods, Qualitative Criteria, Appraisal.

Introduction

Qualitative approaches to evaluation are a powerful tool for analyzing processes in projects, programs and policies. In addition to formative evaluation, qualitative approaches are increasingly playing a role in summative evaluation, where their particular strength lies in uncovering complex causal relationships (Maxwell, 2020). Qualitative evaluation is also increasingly being used in the context of evidence-based policy. However, although the potential of qualitative evaluation designs is evident, their use in evaluation often falls behind quantitative evaluation designs. Reasons for this include the fact that the underlying information is not considered accurate, the credibility of the results is questioned, or that qualitative designs are assumed to be more suitable for small and local evaluations (Henry, 2015).

Moreover, while qualitative results sometimes find their way into policy implementation as influential individual studies, they tend to be more influential when summarized in qualitative syntheses. For this, however, they must rely on common and explicitly reported standards. Carroll et al. (2012), for instance, argue that the explicit reporting of standards is a good measure for the methodological quality of a study. In a sensitivity analysis, they investigate whether the exclusion of insufficiently reported studies from a synthesis influences the conclusions. They conclude that inadequately reported studies made hardly any contribution to the overall result and therefore recommend their exclusion. While such calls remain controversial (Verhage and Boels, 2017), the real-world influence of qualitative evaluations depends on whether they report their results in an appropriate form.

Therefore, the acceptance of qualitative approaches depends on the extent to which they communicate credibility techniques in the production and analysis of empirical evidence (Anastas, 2004; Liao and Hitchcock, 2018). Handbooks of professional organizations have long demanded quality of inference and empirical evidence. Examples are the American Evaluation Association, the Canadian Evaluation Society (Yarbrough et al., 2011) or the various national evaluation associations in Europe (i.e. European Evaluation Society). Criteria for credibility of qualitative evaluation thus exist and the primary question is whether they are also applied and reported in evaluations (Macklin and Gullickson, 2022). The methodological diversity of qualitative evaluations, however, comes with the additional challenge of how to synthesize findings based on heterogeneous methods with heterogeneous purposes (Lawarée et al., 2020). Qualitative approaches comprise different ontological and epistemological positions. As Patton (2015, p. 164) puts it: “Qualitative inquiry is not a single, monolithic approach to research and evaluation.”

In this contribution, we are therefore interested in two related research questions in this context. We are interested in how far qualitative evaluations follow methodological recommendations mentioned in the literature on qualitative research methods, and how they report these standards. A common foundation of qualitative approaches and common practices of reporting thus plays a role in practical and academic discussions of evaluation, especially when we consider the increasing prestige of evidence-based policy (Buckley et al., 2020;

Sanderson, 2002). The findings of qualitative evaluations, in turn, depend on the approaches and methods that are common and accepted in the field. Therefore, we also document the methods employed, instruments used and empirical strategies proposed, as the reported quality standards relate to these. In addition to documenting evaluation methods as they are published, this also helps us to recognize patterns across different methods.

In the following, we analyze in several steps qualitative evaluations published in scientific journals in the period from 2015 to 2019. First, we draw a sample of 1070 articles, which we examine by simple content analysis to show the variety of qualitative approaches and methods used. In a second step, we draw a sample of 50 articles, which we code manually. We follow established reporting and quality criteria for qualitative research. As expected, the results show a broad spectrum of qualitative evaluation approaches – similar to the large sample. Case studies are the most common evaluation design. Stakeholder analysis and community analysis are two of the most important categories of evaluation, while interviews tend to be the most important source of information. When we shift to the reporting of quality criteria, no uniform practice seems yet to emerge. For instance, relatively few articles talk about the role the evaluator plays in the evaluation, let alone a deeper reflection on positionality, and the role of the evaluator in the evaluation context. While this is understandable from a practical point of view, it constitutes a lost opportunity for a more systematic inquiry. We also find few systematic differences between methods. For instance, case studies are not much more likely to report transferability than other types of evaluation designs despite the fact that case studies would naturally lend themselves for such discussions.

In the following section, we discuss criteria and reporting standards in qualitative research and evaluations. In the third section, we explain our strategy regarding the selection and coding of our smaller sample. The fourth section presents the results. In the final section, we discuss the implications of the results.

I. Methods and reporting practices of evaluation

The field of qualitative methods of evaluation is diverse, so our contribution can hardly do justice to all varieties and purposes of evaluation, but there are excellent overviews for qualitative evaluation methods (Bamberger et al., 2009; Befani, 2020; Donaldson and Lipsey, 2006; Donaldson and Scriven, 2003; Maldonado Trujillo and Pérez Yarahuán, 2015; Patton, 2015). We will talk more about the diversity of methods in our section on how we coded the articles, but one source of that diversity is worth highlighting. A major distinction lies in the epistemology on which qualitative methods are built. While many qualitative approaches such as process tracing often follow a positivist logic, others combine insights from positivist and constructivist logics (e.g. Pawson and Tilley, 1997) or are firmly based on interpretative-hermeneutic perspectives (Henry and Mark, 2003).

Perhaps as a result of these different theoretical traditions, few studies have mapped the diversity of the field. There are, however, excellent textbooks both on evaluation research (Patton, 2015; Shaw, 1999) as well as qualitative research methods (Creswell, 2007; Denzin

and Lincoln, 2018; Maxwell, 2013) which provide tools for categorizing different approaches. There is also a vibrant discussion of research on evaluation picking up many of the themes we develop below (Apgar et al., 2024; Azzam, 2011; Christie and Fleischer, 2010; Coryn et al., 2017; Lawarée et al., 2020; Leininger and Schiller, 2024; Miller, 2010; Smith, 1993; Teasdale, 2021; Teasdale et al., 2023).

Against the background of methodological diversity in qualitative evaluation there are calls for professionalization (Picciotto, 2011; United Nations Evaluation Group, 2016), and for common reporting practices in evaluation reports (Carroll et al., 2012; Miller, 2010; Montrosse-Moorhead and Griffith, 2017). While several of these standards already exist, converging on common labelling them remains a challenge. There are multiple appraisal tools for various subject areas (Majid and Vanstone, 2018). Examples of subject-specific standards are the Consolidated Standards of Reporting Trials (CONSORT), which define reporting criteria for RCTs in the medical field (CONSORT Group et al., 2010), the Consolidated Criteria for Reporting Qualitative Research (COREQ) (Tong et al., 2007), which are specific for studies that use interviews and focus groups or the Qualitative Research Guidelines (QRG) (Wu et al., 2016), mainly serve as author guidelines for the Journal of the Society of Social Work and Research. In the rest of this section, we briefly look at the difficulties of developing common quality criteria for qualitative social research and evaluation and look at existing catalogues for common standards in the field of qualitative evaluation.

In the field of evaluation practice, Macklin and Gullickson (2022) show the broad spectrum of conceptualization of the term “validity”. Depending on the underlying research paradigm, validity can be conceptualized as internal or external validity, or as translational or interpretive validity, to name just a few approaches. Validity can also be conceptualized more narrowly in terms of research design, data sources and sampling as well as measurement, or thought of more broadly, which would also include preparation, reporting of results and evaluation use. Here too, scholars find a great diversity of concepts in their synthesis of the corresponding literature.

If we want to examine the quality criteria used in qualitative evaluations, we cannot work with the criteria of the positivist approach alone. Criteria exported from quantitative positivist do not suffice to account for most problems in qualitative evaluation research. However, qualitative researchers have proposed alternative criteria to those used in positivist hypothetico-deductive research (Jacobs et al., 2021: 186–187). One example is “trustworthiness” (Guba and Lincoln, 1989; Lincoln, 2005; Schwartz-Shea and Yanow, 2012: 91–114) of the research results instead of validity, reliability and replicability. The methodological literature offers a whole range of benchmarks in this regard, some of which are specific to certain approaches (cf. Creswell, 2007, p. 203). These criteria cannot be directly translated into positivist terminology, but they can be thought of as functional analogies.

Taking inspiration from Apgar et al. (2024) and others, the following grouped characteristics are crucial for the credibility or trustworthiness of qualitative research as a minimum requirement: Reflexivity, confirmability and transferability.

1) Reflexivity, and, relatedly, positionality, is a fundamental stance of qualitative research (Alvesson and Sköldberg, 2018; Schwartz-Shea and Yanow, 2012: 99–104). Qualitative research acknowledges the impact of the researcher on the participants; researchers are part of the field, so to speak, rather than outside observers. This point of view is important for evaluation, especially for participatory forms of evaluation. It is also important for readers of evaluation research to better understand the role the evaluator plays in the evaluation.

2) Another criterion is transparency or confirmability. Since qualitative methods have to reconcile various sensitive issues, such as protecting sources, building trust, and honoring contextual information, transparency of research is secured by confirmable procedures, such as "member checking" (Schwartz-Shea and Yanow, 2012: 106), a detailed description of the process of analysis, coding procedures and inter-coder agreement (Creswell, 2007: 209–210). Similar to positivist approaches, however, we also consider – in line with Carroll et al. (2012) – the justification of the choice of methods and selection of data sources to be important for a transparent approach.

3) A final criterium is the transferability of results (Schwartz-Shea and Yanow, 2012: 47–48), which is often not the central aim of qualitative research and evaluation. Still, generalizability is an important issue, as evaluation research wants to highlight aspects that are important beyond the specific intervention. In synthetic or systematic reviews this criterium plays a huge role. We are therefore interested in the transferability of the results to other cases with a similar context (Guba and Lincoln, 1989; Lincoln, 2005). To give just one example: a "thick description" of the case under evaluation might help to transfer findings of the case into a wider field of similar cases (Schwartz-Shea and Yanow, 2012: 145) or also a discussion under which contextual conditions the results are transferable. All in all, we assume that reflexivity, confirmability and transferability ensure the trustworthiness of qualitative evaluations as minimum requirements.

Table 1. Overview for Qualitative Criteria in Social Sciences.

Quantitative	Qualitative
Objectivity/ Inter-subjectivity	Reflexivity/ Positionality/(In-)Dependence
Validity, Reliability, Replicability	Transparency/ Confirmability
Generalizability	Transferability

Note: Own compilation on basis of the sources mentioned above.

Table 1 gives our stylized overview over the methodological debate on quality criteria[1]. Note that we do not claim that those standards should only relate to quantitative or qualitative approaches exclusively. We also do not want to draw strong analogies between a standard such as objectivity on the quantitative side, and reflexivity on the qualitative side. With these clarifications in mind, Table 1 (right-hand side) will inform most of our coding for the quality criteria in the rest of the paper.

II. Selection of articles and coding procedure

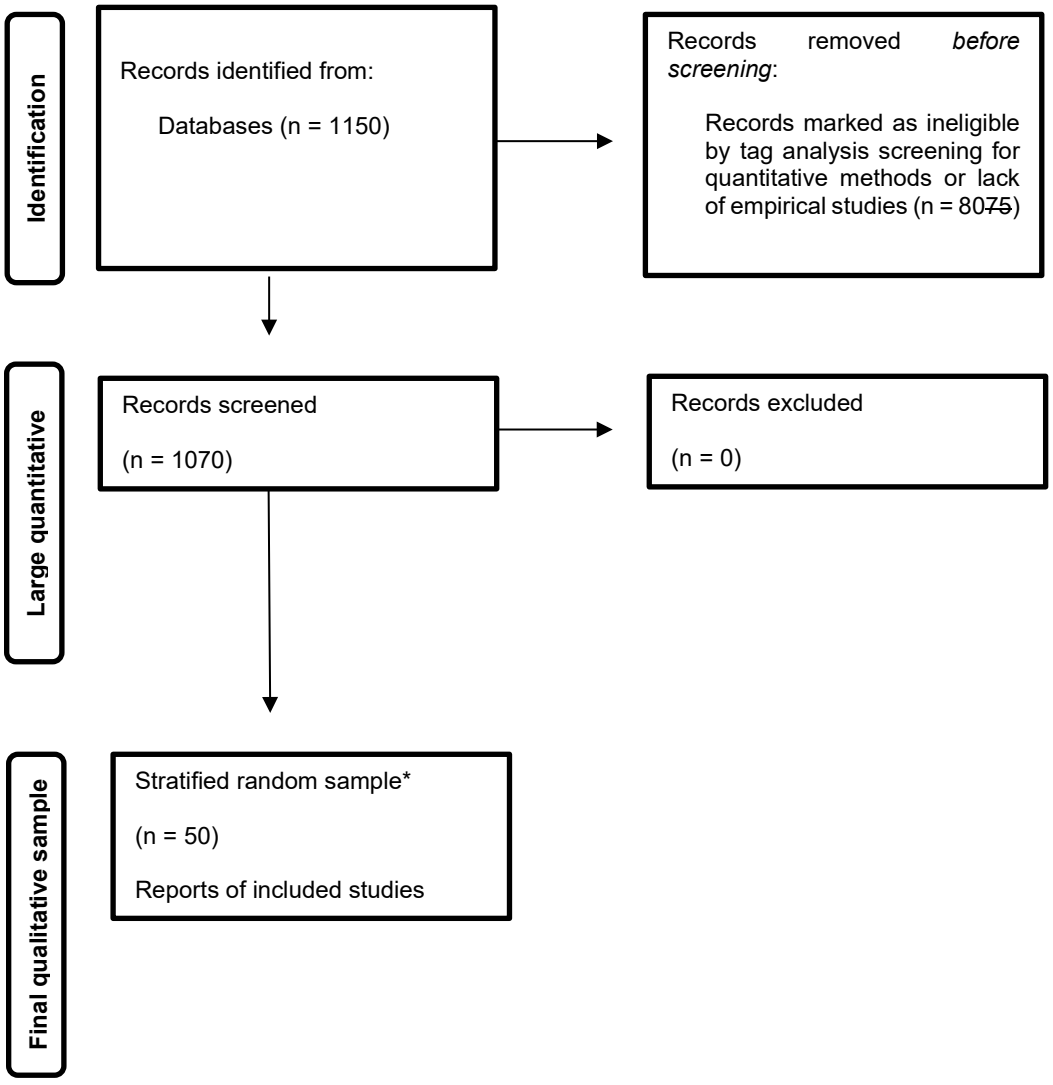
Qualitative researchers highlight the need to be transparent about researchers' own positions as well as initial failures and a learning in the research process (Jacobs et al., 2021). Therefore, we start with some reflections on our own independence and positionality. We are both academic researchers with practical experience in evaluation, but not involved in any organization or ongoing evaluation activity. In particular, we have no direct relationship to any of the coded articles analyzed below. Given our personal research experience, we may both have a bias towards quantitative methods of evaluation and the positivist paradigm, as our own research often uses quantitative studies. We thus try to be as transparent about the following research methodology and our trial-and-error approach to making sense of the information.

Our selection of articles has two aims. First, it should cover a broad range of relatively comparable qualitative evaluations, and second, it should be accessible to a broad readership. We decided to only look at articles published in peer-reviewed journals, as these are read by a broad readership, are a relatively homogeneous group, and often focus on issues of methodology. Similar to Christie and Fleischer (2010), Coryn et al. (2017) and Teasdale et al. (2023), we focus on major, general journals of evaluations. This also implies that we do not include grey literature which might sometimes have more methodological appendices and details.

[1] See Macklin and Gullickson (2022) for a more comprehensive list on validity criteria.

To select the journals, we first looked at the full list offered in Social Sciences Citation Index, especially those listed as "social science interdisciplinary" and "evaluation" journals. We have avoided specific journals that only focus on evaluating certain areas such as health, education etc. Our final list includes American Journal of Evaluation, Canadian Journal of Program Evaluation, Evaluation, Evaluation and Program Planning and Evaluation Review. For practical reasons, we have opted for a period of five years (2015–19), resulting in a total of some 1150 articles (see Appendix 1).

Figure 1. PRISMA flow chart for identification of studies via databases and registers. Flow diagram adopted from Page et al. (2021).



* Based on new random sample stratified by year.

The flow diagram in Figure 1 illustrates our approach. In our first attempt, we checked all articles for qualitative evaluations, using tags (and keywords) provided by the journals as a first filter. In particular, we excluded studies that entailed tags associated with purely quantitative analysis or theoretical contributions without reference to any sort of cases or empirics. This first selection resulted in 1070 articles, the total number of qualitative articles in the selected journals for the period under review. With these articles, we did some simple keyword searches to get a feeling for the overall collection of articles. Although this crude and simplistic quantitative content analysis leaves much to be desired for (cf. below) and was not our main aim, it does give some indication of the relationships between all articles and those finally selected and coded.

We tested our initial coding scheme with a random sample of 95 papers. We coded these articles according to a simple coding scheme that gives us a first overview of the distribution of the articles in relation to the categories listed in Appendix 2. We coded 54 articles separately, and 41 articles jointly. We then checked the jointly coded articles for inter-coder reliability, which resulted low at 52.2 percent (Krippendorff, 2011: 211–256). After discussing and correcting the discrepancies in our coding strategies, we decided to adapt our coding strategy and to use a two-stage deliberative coding procedure (e.g. Hak and Bernts, 1996; Saldaña, 2021).

In our final approach we started with a new stratified random selection of 10 articles from each year to reach our final sample of 50 articles from 2015–19. Stratification means that we draw a random sample from a specific group (stratum), using the year of publication as the stratum to obtain a better-balanced final sample. The random selection of articles mainly serves to minimize our own bias in terms of topics and methods. In the first round, we coded the articles separately. In the second round, we discussed each article and made adjustments to our coding. In very few cases, we could not agree on a common coding because we either had different impressions or the articles were not clear enough. However, for the vast majority of the codings, we agreed after the second stage, bringing the final intercoder reliability to 96.4 per cent (Appendix 3). This high result is not surprising given the deliberative approach. It simply documents the change in our own research strategy and gives some indication of where problems may have arisen in our coding strategy[2].

[2] We also coded a random sample of 20 papers from 2020-2022 as a control. The results are similar to those of the sample of 50 papers from 2015-2019 in Figure 2.

III. Results for Sample Characteristics and Types of Methods

Before looking at the results of the small sample of 50 articles, we present some comparative information for the initial, larger sample of 1070 articles and compare it to the smaller, final sample (see list of articles in Appendix 9). The comparison should be treated with caution, as the coding of the large sample was done by simple keyword search. As is well known (e.g. Krippendorff, 2011), simple (lexicographic) keyword searches have validity issues. For instance, our keyword searches resulted in a large number of missing observations (see below). Nevertheless, comparing the keyword searches with our deliberative coding reveals interesting cross-cutting forms of evidence and improves the transferability of the overall findings. When we compare the distribution of the studies' evaluands in terms of sectors, we see that the sectoral distribution is relatively similar between the two samples (table 2). This suggests that the smaller sample of 50 articles is not very different from all articles between 2015 and 2019.

Table 2. Sectoral Distribution of Coded Articles

	50 articles		all articles	
Sector	frequency	percent	Fre-quency	percent
Development	6	8	115	11
Education	12	17	97	9
Social & Labour	8	11	137	13
Economic	4	6	32	3
Health	30	42	258	24
Administration	5	7	41	4
Infrastructure	1	1	8	1
Crime/Security	1	1	12	1
Other	4	3	358	34

3.1. Results for Qualitative Methods, Designs and Sources

We briefly document the enormous diversity of qualitative evaluation methods and their contexts. As discussed above, we have collected information on evaluation context, and the types of evaluations methods. We highlight some of those findings for the small sample of 50 articles. We start by looking at major categories of evaluations for which we took inspiration from Kusters (2011), Mathison (2005), Newcomer et al. (2015), Patton (2015), Shaw (1999) and Shaw et al. (2006). Clear demarcations between these categories are difficult. We also realized this in our own coding process. For this reason, Table 3 shows two types of frequencies. One shows the first codes we gave. The idea behind the first codes is that it is the most intuitive choice of both coders, the category that first came to mind while reading the article. We contrast this with all codes given, in those cases where we allowed for multiple codes. Evaluation categories are such a case of multiple codings. Regardless of using first or all codes, we find that stakeholder analysis, in a broad sense of the word is, by far, the most

frequent category. The next categories are developmental and community-based evaluations, followed by mixed methods and evaluations that contain some form of theory of change or logic model.

Some of these categories co-occur with each other relatively frequently (Appendix 4). For instance, community and participatory approaches often coincide, as do participatory approaches and stakeholder analysis. It is also not surprising that realist evaluations often contain a theory of change. Others are more orthogonal, i.e. they do not co-occur so often with other categories, especially the less frequent categories. Here are some examples of less frequent combinations: (Millett et al., 2016) who combine theory of change with developmental evaluation; (Suiter, 2017) who combines realist and community evaluation; Chen (2017) who combines stakeholder analysis with a quantitative approach in an importance-performance analysis; Frye et al. (2017) who combine community analysis, realist evaluation and theory of change with an quasi-experimental approach and Koleros et al. (2016) who use a mixed-methods approach with community and contribution analysis and theory of change. In general, we note that the different categories and approaches are very diverse, but some categories are clearly more dominant (see table 3). We also need to emphasize that larger categories such as stakeholder analysis hide a great deal of heterogeneity within the category (see below).

Table 3. Evaluation Categories.

	1st codes		All codes	
	Fre-quency	Percent	Fre-quency	Percent
Stakeholder analysis	17	41	23	32
Community approach	5	12	6	8
Realist evaluation	3	7	3	4
Contribution analysis	1	2	2	3
Participatory approach	2	5	9	13
Developmental evaluation	4	10	8	11
Process tracing	1	2	1	1
Mixed methods	3	7	7	10
Theory of change/ logic model	3	7	9	13
Network analysis	2	5	2	3
Experimental	0	0	2	3
Total	41	100	72	100

The diversity of qualitative methods is also evident in Table 4, which lists the different empirical methods of information gathering that we have coded following Patton (2015). All articles contained interpretable information on the main empirical sources used. The table again shows both first codes and all codes. The most common category here is “interviews”. This comes at no great surprise since stakeholder, community and participatory approaches dominate the previous table. Reporting on the exact nature of these interviews varied greatly from article to article, so it is difficult to make a summary assessment. However, there are clearly major differences in interview methods in the sampled articles.

Documents in a broad sense are the second most important category, but here too the heterogeneity is considerable. Often “documents” refers to official documents (such as policies, laws, regulations), but the category also contains field notes, protocols and other types of written records produced during the evaluation. Some categories were less common than we

expected. For instance, anthropological methods such as direct observation are rare. Occasionally we also found it difficult to differentiate between different categories – for instance, focus groups and participatory workshops were sometimes indistinguishable.

Table 4. Data Sources.

	1st codes		All codes	
	Frequency	Percent	Frequency	Percent
Survey	11	22	12	12
Interviews	20	40	28	29
Focus groups	3	6	18	18
Documents	15	30	26	27
Participatory workshops	0	0	10	10
Direct observations	1	2	4	4
Total	50	100	98	100

The appendices report further results on the different types of evaluation methods. The most frequent research design we could identify was the case study method (Appendix 5). Similar to "stakeholder analysis", "case study" is a very broad category. Very few articles gave explicit definitions of the term "case" or discussed in detail issues related to the design of case study research. It often remained implicit, what the universe of cases should be and how the case studied relates to this universe.

Most articles contained some form of program or project evaluation, but there were also many studies that focused on research on evaluation. Explicit meta-studies (meta-narratives, meta-analysis etc.) were comparatively rare, as we excluded most quantitative evaluations (Appendix 6). In terms of research paradigm, three quarters of all articles followed some form of the positivist paradigm. Constructivist and interpretative approaches were relatively rare (Appendix 7). Similarly, "causal analysis" was the most frequent category, but closely followed by descriptive and explorative studies (Appendix 8).

We see in this overview that three approaches and three methods are predominantly chosen to generate qualitative evaluation evidence. Stakeholder analysis, community approach and developmental evaluation account for 63 percent of the main approaches chosen. The most important research design is case studies. Data is collected primarily through interviews, documents and surveys (92 percent). This is interesting insofar as qualitative evaluations have a

broad spectrum of approaches and methods at their disposal. In practice, however, they focus on far fewer techniques, almost as if there were an informal consensus on how to achieve qualitative results. At the same time, some studies take a more holistic view by combining several approaches, designs and data sources. Using this information of the types of methods, we can now turn to the reporting criteria to see whether there are differences across those methods, broadly defined.

3.2. Reporting of Quality Criteria

Montrosse-Moorhead and Griffith (2017) developed a comprehensive catalogue of criteria that provides a common quality basis across disciplines and methods. Their Checklist for Evaluation-Specific Standards (CHESS) takes into account a full range of criteria for all types of evaluation approaches. In this section, we use CHESS as an inspiration and combine it with the credibility criteria listed in table 1 to code our articles.

Let us begin with perhaps the most important category – evaluator independence, which can be related to questions of reflexivity and positionality. Although explicit discussion of independence was rare, we tried as best as we could to code information about the authors of the articles to assess their role in the evaluation. We have distinguished three cases: In the first case, the authors are both evaluators and policy makers. By the latter, we mean that they are also responsible for the intervention itself. This was the case for almost 30 percent of the articles. This form of (in)dependence prevails in participatory or developmental evaluations, for example, when representatives of health organizations are part of the evaluation team and also part of the team that wrote the research article. More common, however, is the situation where the authors are also the evaluators but are not responsible for the intervention itself, which is the case in 56 percent of the articles. This is perhaps the classic case of a formally independent evaluation.

Of course, we know very little about the detailed social background of the evaluators and how close they are to those responsible for the intervention. For us independence was expressed only by method of exclusion: the authors did not appear to be involved in the original design of the intervention. Finally, in only 10 percent of the articles were the authors different from the evaluators and policy makers. This was the case when we coded meta-studies, i.e. authors who reviewed other people's evaluations. Transparency with regard to independence helps readers of the evaluation study to assess the quality of the results more accurately, as Sturges (2015) shows:

> "*I became involved in College Now (CN) when I was recruited during the program's third year by the firm hired to perform its evaluation. One of my first evaluations, I oversaw classroom observations, interviews with students and teachers, and AP course plans and materials. At the same time, I was conducting a qualitative study on reform at two CN schools. [...] The simultaneity of projects provided me a unique insider–outsider perspective on CN and helped me reflect on my responsibilities as a junior evaluator.*" (p.462)

We also looked more closely at reflexivity and positionality. In our coding, "reflexivity and positionality" was present when authors explicitly described their own role as well as potential biases, subjective interpretation, and reflexive relationships. Figure 2 shows the percentage frequencies of simple yes and no dummy questions when we considered a quality criterion to be explicitly discussed (1) or otherwise (0). "Reflexivity" is the first bar in Figure 2, showing that only a small percentage of articles address reflexivity and positionality. Similarly, a keyword search for reflexivity in the large sample of 1070 only gives 2 direct hits. And yet, there are exemplary accounts of reflexivity, such as Siebert and Myles (2019) deliberate approach to the implications of their own roles as external evaluators:

> "*Our initial failure to reconstruct a programme theory that provided an accurate representation of the programme made us reflect on our role as evaluators. How much direction do we as evaluators provide when this becomes apparent? Should we reconstruct the theory for the programme as we perceive it, or should we focus on using methods that will facilitate a process that will enable stakeholders to do so themselves? […] It was these questions and reflections that made us decide to find a way to remove misrepresentation and be assured that the process of reconstructing the programme theory was both transparent and representative of stakeholders'collective logic*." (p.471)

It is obvious that not all types of evaluation methods require a deeper engagement with reflexivity and positionality. Nonetheless, we think it is worth considering whether mainstreaming such considerations might not be an important aspect of all evaluation research. Qualitative researchers are very sensitive to such issues and could only strengthen their own contributions by being open about their positionality. Although some might see this to as weakening their own evidence,[3] such a concern should be irrelevant in an academic context. Instead, if reflections on reflexivity and positionality become part of the reporting routine, this could greatly facilitate readers' understanding of the evaluation context and evaluation research.

The next major quality criterium is transferability (Figure 2, first row). According to our coding, nearly 50 percent of all articles discuss the transferability of findings to other cases, domains, fields, or interventions. What exactly counts as transferability is, of course, context-dependent, but it is enormously helpful for a reader of such articles to know the author's assessment of which parts of the general lessons would be transferable to other applications. Note that we coded transferability as "1" even when the authors explicitly mention that generalizations from the case are difficult or even impossible for specific reasons.

Often it is useful to dismiss the idea of generalizability, but in most cases explicit discussions are helpful, especially when authors refer to their work as a "case study". Against our expectations those studies for which we coded the research design category as 'case study' did not have a significantly higher proportion of discussions of transferability. In about 50 percent of all case studies the transferability of findings was discussed (see below). In our opinion, this is a missed opportunity, because an open discussion about what the authors think in this regard

[3] Background interview with a professional evaluator at a major institution evaluating development projects. Name and details must remain confidential.

would be very helpful. It would give us the means to synthesize the results and create systematic reviews based on many cases. Nevertheless, there are interesting examples of how transferability of results can be communicated. For example, the role of context in the transferability of results is also highlighted by Nielsen et al. (2019):

> "*A strength of this study is the focus on the provider perspective of the implementation processes and the contextual descriptions of a successful school-based programme tripling the amount of PE. (…) This potentially increases practitioners and decision-makers ability to assess the programme in relation to their individual context (…) – ultimately strengthening the transferability of the programme and the strategies used to secure the implementation of additional PE or PA in a school context."* (p. 7)

The importance of confirmability as a quality criterion is highlighted in Table 1. Such a criterion is, of course, not easy to code. One thing we coded directly was whether the article explicitly mentioned data repositories, appendices, or other further information about the empirical database or methodological appendices. Here again, we coded this criterion as "1" even if the authors explicitly stated that the information could not be shared due to data privacy issues. Very few studies explicitly stated such concerns about confirmability. Given that such direct reporting on transparent or confirmable reporting is relatively rare we also looked for instances, where authors motivated the choice or thoroughly described their methodological approach. Here we found several examples, such as in Kokko and Lagerkvist (2017):

> *"All interviews were transcribed for analysis, which involved open coding of narrative descriptions according to the grounded theory generation procedure described by Glaser and Strauss (1967), and development of thematic categories and abstraction of conceptual metaphors to categories. The qualitative data organization software Atlas.ti Version 7.5.7 was used for organization of coding and categorization. When coding, especially in the construct elicitation step, the aim was to make broad enough categories of meaning for the elements of the ladders (A-C-V) in order to obtain links identified by more than one participant, without losing the relationships between the elements and not focusing on the elements themselves (…). Applicable codes were created for pair–construct relationships, such as health and sickness, save or spend money, and sustaining or difficulty in sustaining daily living, (…)."* (p. 211)

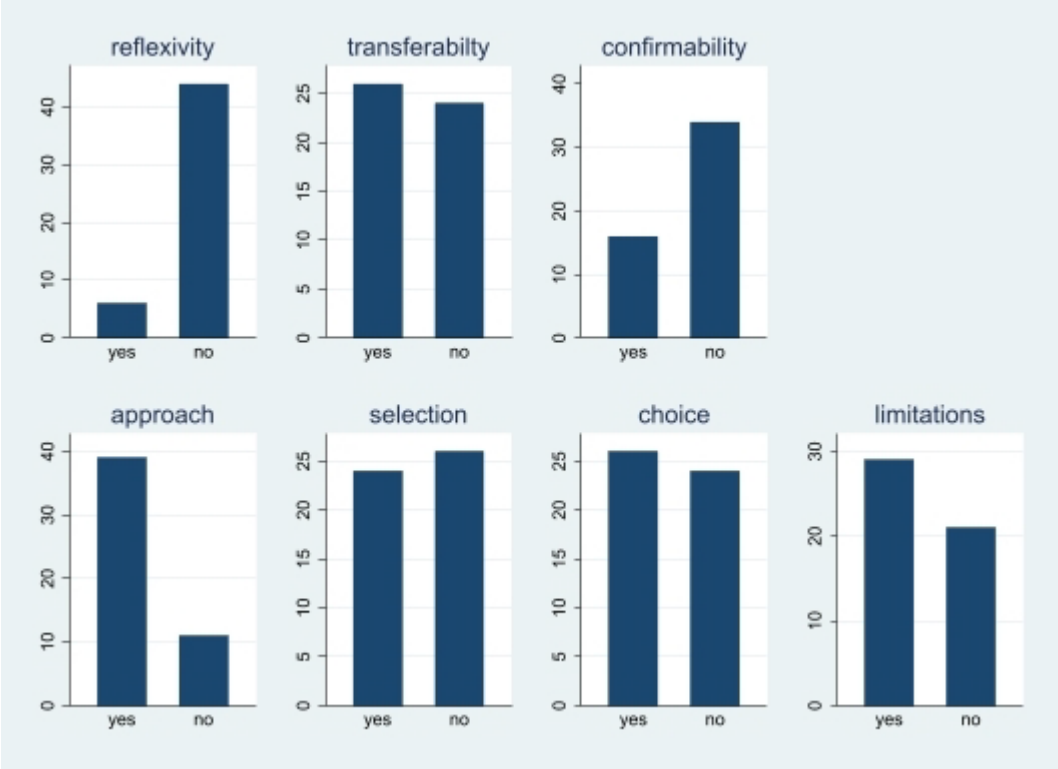
Part of confirmability, then, is whether the articles justify the choice of the approach, the choice of empirical source, and the selection of observations (if applicable). Here, the articles provided much more detailed information (see second row of Figure 2). Most articles implicitly or explicitly discussed why a particular approach ("approach" in Figure 2) was used – say realist evaluation or contribution analysis. About half of the articles also justified the empirical sources ("choice") and observations ("selection"). Although not all studies lend themselves to such discussions, it is interesting to learn why an evaluation mainly relies on, for instance, interviews rather than focus groups or why it selects certain types of stakeholders but not others. Sometimes a brief justification for the choice of method is sufficient to help readers:

> "*We also recognised that a consensus view of the programme needed to be achieved for things to progress. To do this, we decided to hold a participatory workshop drawing on the principles of Leeuw's (2003) strategic assessment approach*." (Siebert and Myles, 2019: 471)

It is interesting to note that there are few systematic differences for the reporting on confirmability across different designs or categories. In other words, it does matter little if we looked at an article using stakeholder analysis or developmental evaluation, or both. When we look at the rationale for specific choices given it is also interesting to note that few articles go into details. Let us take the example of stakeholder analysis, to most frequent category identified above. While it is difficult to divide stakeholder analysis into different groups, clearly qualitative evaluators can have different types of stakeholders in mind. To give but two examples: In a participatory approach, stakeholders include all those affected by an intervention (target group), whereas an evaluator guided by the concept of veto players would focus only on stakeholders who are powerful or institutionally relevant actors (see e.g. Jepsen and Eskerod, 2009; Reed et al., 2009). In practice, we found few such discussions in the articles, but they might be helpful to understand more systematically which type of stakeholder analysis was done.

A final criterion is the simple reporting of limitations. Many articles – rather than reporting any explicit standards – deal with limitations explicitly and thereby reveal potential weaknesses or shortcoming in a transparent way. In this sense, this regular practice is universal and includes all types of evaluation methods. Figure 2 shows that 60% of all studies talk about their own (methodological) limitations. In the remaining cases, it might make sense not to mention the limitations, but it might again be a missed opportunity.

Figure 2. Frequencies of Standards Reported.



Note: Own graph on basis of coding for 50 articles.

3.3. Limitations of the study

Our analysis of methods and reporting criteria thus leads to mixed results. On the one hand, we find great diversity of methods and great examples in dealing with the reporting criteria we have selected. On the other hand, evaluation practice tends to concentrate on a few approaches and methods and only partially reports the criteria. We also did not find huge differences in reporting standards across methods and designs. One reason for this could be that the authors publish the articles in scientific journals to highlight points of interest that arise from the evaluation, but which are only excerpts from the underlying evaluations and their findings. Often such information can be found in other documents, e.g. in the detailed evaluation reports, to which we did not have access. Although we did our best to track down such documents where they were mentioned, we cannot really verify this in our analysis.

Other coding difficulties apply to our own analysis. We went through our codebook and our codings several times and yet we may have missed passages in individual articles. We are also aware that our categorizations can be challenged. For example, it can be debated whether theoretical tools such as a "Theory of Change" are part of the same group as "Stakeholder Analysis" or "Developmental Evaluation". The sources we have chosen for categorizations, such as handbooks and textbooks, can be criticized. However, we hope to use intuitive categorizations that will stimulate discussion about research on evaluation methods.

In addition, Liao and Hitchcock (2018) also note the possibility that journal editorial practices could lead to limited descriptions of procedures in evaluations, especially given space constraints. Indeed, we find some, if limited variation for the five journals under inspection (results available on request). While our sample is too small to explore this hypothesis in more detail, we believe that differences in editorial policies could be an interesting area of further research.

To repeat, our small sample selection does not claim to be representative. The population from which a random sample should be drawn is not evident given the large number of publication outlets. Therefore, we sought to cover a broad range of evaluation areas and to minimize our own bias. We plan to follow-up our analysis with a more thorough quantitative content analysis of a larger set of evaluations. This might also allow us to find more nuanced differences in reporting standards across methods and designs. Still, we think that our findings have some degree of transferability, both to the larger sample as well as other types of publications reporting evaluation results.

Conclusion

Our investigation is motivated by the question of how qualitative evaluations can strengthen the visibility and relevance of their results. A key measure is to follow reporting standards on evaluation design that allow consumers of the study to assess the quality of the results. The diversity of qualitative evaluation approaches and methods, which are rooted in different ontologies and epistemologies, naturally makes it difficult to establish uniform reporting standards. In addition, the relevance of the results also depends on which approaches and methods are primarily used in qualitative evaluations.

To understand methodological and reporting practices in more detail, we coded and analyzed a sample of 50 qualitative evaluation studies. We find that despite considerable diversity in terms of approaches and methods – which is the strength of qualitative evaluations – only a few approaches and methods dominate in practice. We have also found that qualitative evaluators often combine different metatheoretical, methodological, and empirical approaches and succeed in making a holistic assessment of the evaluation problem.

Our analysis of reporting practice reveals a mixed picture. We apply a minimal set of criteria covering the domains of reflexivity, confirmability, and transferability. We find many good examples of reporting these criteria, but many evaluation studies do not allow us to draw conclusions about their underlying quality standards. There is room for improvement in the documentation and reporting of qualitative contributions to evaluations, particularly in two respects. First, while it is not always easy to do so, some more explicit conceptualizations of key aspects of the methodology would help readers assess the material. As mentioned earlier, stakeholder analysis is a good example. For qualitative evaluation researchers, it is often intuitive who the key stakeholders are and how to involve them. For readers, this is not always the case. There are different variations of stakeholder analysis – from participatory approaches

to those that emphasize the role of powerful veto players – and the difference matters for evaluation results.

Similarly, an explicit discussion of other recurring methodological tools would be helpful, such as the types of interviewing techniques or the types and applications of case study analysis. For example, why do evaluators use the term “case study”, with what notion of a “case”, and what might other, comparable cases be? We believe that a more detailed account of these conceptual issues would go a long way toward building bridges and synthesizing findings or knowing when not to synthesize them (Carroll et al., 2012).

A second, similarly important area for improvement is in reporting of criteria. We believe that the particular nature of evaluation merits a much more explicit discussion of crucial aspects of the process. What is the role of evaluators in evaluation and intervention? To what extent do they reflect on their own position and their contribution (or perhaps destruction) of the intervention? Qualitative researchers are inherently attuned to these questions, so it would be a strength, not a weakness, to engage more deeply with these issues. We think this is also necessary from an ethical perspective, as evaluators are very close to the target population and their role in the evaluation process. The categories we used to analyze the articles are only one of several possibilities. This is one of the limitations that may affect the transferability of our results. However, we hope that we have made our own methodology sufficiently transparent to elicit justified criticism. With our analysis, we therefore aim to initiate a discussion about methods and reporting standards for qualitative evaluations, but not to provide definitive answers.

Bibliography

ALVESSON M and SKÖLDBERG K (2018) *Reflexive Methodology: New Vistas for Qualitative Research*. 3rd edition. Los Angeles: Sage.

ANASTAS JW (2004) Quality in qualitative evaluation: Issues and possible answers. *Research on Social Work Practice* 14(1): 57–65.

APGAR M, BRADBURN H, ROHRBACH L, et al. (2024) Rethinking rigour to embrace complexity in peacebuilding evaluation. *Evaluation*: 13563890241232405.

AZZAM T (2011) Evaluator characteristics and methodological choice. *American Journal of Evaluation* 32(3): 376–391.

BAMBERGER M, RAO V and WOOLCOCK MJV (2009) *Using Mixed Methods in Monitoring and Evaluation: Experiences from International Development*. Manchester: Brooks World Poverty Institute, University of Manchester.

BEFANI B (2020) Choosing appropriate evaluation methods. A tool for assessment & selection. Guildford: Centre for the Evaluation of Complexity Across the Nexus.

BUCKLEY PR, FAGAN AA, PAMPEL FC, et al. (2020) Making evidence-based interventions relevant for users: A comparison of requirements for dissemination readiness across program registries. *Evaluation Review* 44(1): 51–83.

CARROLL C, BOOTH A and LLOYD-JONES M (2012) Should we exclude inadequately reported studies from qualitative systematic reviews? An evaluation of sensitivity analyses in two case study reviews. *Qualitative Health Research* 22(10): 1425–1434.

CHEN KH-J (2017) Contextual influence on evaluation capacity building in a rapidly changing environment under new governmental policies. *Evaluation and Program Planning* 65: 1–11.

CHRISTIE CA and FLEISCHER DN (2010) Insight into evaluation practice: A content analysis of designs and methods used in evaluation studies published in North American evaluation-focused journals. *American Journal of Evaluation* 31(3): 326–346.

CONSORT Group, SCHULZ KF, ALTMAN DG, et al. (2010) CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *BMC Medicine* 8(1): 18.

CORYN CLS, WILSON LN, WESTINE CD, et al. (2017) A decade of research on evaluation: A systematic review of research on evaluation published between 2005 and 2014. *American Journal of Evaluation* 38(3): 329–347.

CRESWELL JW (2007) *Qualitative Inquiry and Research Design: Choosing among Five Approaches*. Los Angeles, CA. London New Dehli Singapore Washington DC: SAGE Publications.

CRESWELL JW and POTH CN (2018) *Qualitative Inquiry & Research Design: Choosing among Five Approaches*. 4th edition. Los Angeles: SAGE Publications.

DENZIN NK and LINCOLN YS (eds) (2018) *The SAGE Handbook of Qualitative Research*. 5th edition. Los Angeles London New Delhi Singapore Washington DC Melbourne: SAGE Publications.

DONALDSON SI and LIPSEY MW (2006) Roles for theory in contemporary evaluation practice: Developing practical knowledge. In: Shaw IF, Greene JC, and Mark MM (eds) *The SAGE Handbook of Evaluation: Policies, Programs and Practices*. London ; Thousand Oaks, CA: SAGE Publications, pp. 56–75.

DONALDSON SI and SCRIVEN M (eds) (2003) *Evaluating Social Programs and Problems: Visions for the New Millennium*. Mahwah, N.J: Lawrence Erlbaum.

FRYE V, PAIGE MQ, GORDON S, et al. (2017) Developing a community-level anti-HIV/AIDS stigma and homophobia intervention in New York City: The project CHHANGE model. *Evaluation and Program Planning* 63: 45–53.

GUBA EG and LINCOLN YS (1989) Judging the quality of fourth generation evaluation. In: Guba EG and Lincoln YS (eds) *Fourth Generation Evaluation*. Newbury Park, CA: SAGE Publications, pp. 228–251.

HAK T and BERNTS T (1996) Coder training: Theoretical training or practical socialization? *Qualitative Sociology* 19(2): 235–257.

HENRY GT (2015) When getting it right matters: The struggle for rigorous evidence of impact and to increase its influence continues. In: Donaldson SI, Christie CA, and Mark MM (eds) *Credible and Actionable Evidence: The Foundation for Rigorous and Influential Evaluations*. SAGE Publications, pp. 65–82.

HENRY GT and MARK MM (2003) Beyond use: Understanding evaluation's influence on attitudes and actions. *American Journal of Evaluation* 24(3): 293–314.

JACOBS AM, BÜTHE T, ARJONA A, et al. (2021) The qualitative transparency deliberations: Insights and implications. *Perspectives on Politics* 19(1): 171–208.

JEPSEN AL and ESKEROD P (2009) Stakeholder analysis in projects: Challenges in using current guidelines in the real world. *International Journal of Project Management* 27(4): 335–343.

KOKKO S and LAGERKVIST CJ (2017) Using Zaltman metaphor elicitation technique to map beneficiaries' experiences and values: A case example from the sanitation sector. *American Journal of Evaluation* 38(2): 205–225.

KOLEROS A, JUPP D, KIRWAN S, et al. (2016) Methodological considerations in evaluating long-term systems change: A case study from eastern Nepal. *American Journal of Evaluation* 37(3): 364–380.

KRIPPENDORFF K (2011) *Content Analysis: An Introduction to Its Methodology*. SAGE Publications.

KUSTERS C (ed.) (2011) *Making Evaluations Matter: A Practical Guide for Evaluators*. Amsterdam: Amsterdam University Press.

LAWARÉE J, JACOB S and OUIMET M (2020) A scoping review of knowledge syntheses in the field of evaluation across four decades of practice. *Evaluation and Program Planning* 79: 101761.

LEININGER J and SCHILLER AV (2024) What works in democracy support? How to fill evidence and usability gaps through evaluation. *Evaluation* 30(1): 7–26.

LIAO H and HITCHCOCK J (2018) Reported credibility techniques in higher education evaluation studies that use qualitative methods: A research synthesis. *Evaluation and Program Planning* 68: 157–165.

LINCOLN YS (2005) Fourth generation evaluation. In: Mathison S (ed.) *Encyclopedia of Evaluation*. Thousand Oaks, CA: SAGE Publications, pp. 162–164.

MACKLIN J and GULLICKSON AM (2022) What does it mean for an evaluation to be 'valid'? A critical synthesis of evaluation literature. *Evaluation and Program Planning* 91: 102056.

MAJID U and VANSTONE M (2018) Appraising qualitative research for evidence syntheses: A compendium of quality appraisal tools. *Qualitative Health Research* 28(13): 2115–2131.

MALDONADO TRUJILLO C and PÉREZ YARAHUÁN G (eds) (2015) *Antología Sobre Evaluación. La Construcción de Una Disciplina*. Carretera México-Toluca: CIDE, Centro de Investigación y Docencia Económicas.

MATHISON S (ed.) (2005) *Encyclopedia of Evaluation*. Thousand Oaks, CA: SAGE Publications.

MAXWELL JA (2013) *Qualitative Research Design: An Interactive Approach*. 3rd edition. 41. Los Angeles London New Delhi: SAGE Publications.

MAXWELL JA (2020) The value of qualitative inquiry for public policy. *Qualitative Inquiry* 26(2): 177–186.

MILLER RL (2010) Developing standards for empirical examinations of evaluation theory. *American Journal of Evaluation* 31(3): 390–399.

MILLETT LS, BEN-DAVID V, JONSON-REID M, et al. (2016) Understanding change among multi-problem families: Learnings from a formative program assessment. *Evaluation and Program Planning* 58: 176–183.

MONTROSSE-MOORHEAD B and GRIFFITH JC (2017) Toward the development of reporting standards for evaluations. *American Journal of Evaluation* 38(4): 577–602.

NEWCOMER KE, HATRY HP and WHOLEY JS (2015) *Handbook of Practical Program Evaluation*. 4th edition. San Francisco: Jossey-Bass & Pfeiffer Imprints, Wiley.

NIELSEN JV, BREDAHL TVG, BUGGE A, et al. (2019) Implementation of a successful long-term school based physical education intervention: Exploring provider and programme characteristics. *Evaluation and Program Planning* 76: 101674.

PAGE MJ, MCKENZIE JE, BOSSUYT PM, et al. (2021) The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*: n71.

PATTON MQ (2015) *Qualitative Research & Evaluation Methods: Integrating Theory and Practice*. 4th edition. Los Angeles London New Delhi Singapore Washington DC: SAGE Publications.

PAWSON R and TILLEY N (1997) *Realistic Evaluation*. London ; Thousand Oaks, CA: SAGE Publications.

PICCIOTTO R (2011) The logic of evaluation professionalism. *Evaluation* 17(2): 165–180.

REED MS, GRAVES A, DANDY N, et al. (2009) Who's in and why? A typology of stakeholder analysis methods for natural resource management. *Journal of Environmental Management* 90(5): 1933–1949.

SALDAÑA J (2021) *The Coding Manual for Qualitative Researchers*. 4th edition. Thousand Oaks, CA: SAGE Publications.

SANDERSON I (2002) Evaluation, policy learning and evidence-based policy making. *Public Administration* 80(1): 1–22.

SCHWARTZ-SHEA P and YANOW D (2012) *Interpretive Research Design: Concepts and Processes*. 1st edition. New York, NY: Routledge.

SHAW I (1999) *Qualitative Evaluation*. SAGE Publications.

SHAW I, GREENE JC and MARK MM (2006) *The SAGE Handbook of Evaluation: Policies, Programs and Practices*. London: SAGE Publications.

SIEBERT P and MYLES P (2019) Eliciting and reconstructing programme theory: An exercise in translating theory into practice. *Evaluation* 25(4): 469–476.

SMITH NL (1993) Improving evaluation theory through the empirical study of evaluation practice. *Evaluation Practice* 14(3): 237–242.

SUITER SV (2017) Community health needs assessment and action planning in seven Dominican bateyes. *Evaluation and Program Planning* 60: 103–111.

TEASDALE RM (2021) Evaluative criteria: An integrated model of domains and sources. *American Journal of Evaluation* 42(3): 354–376.

TEASDALE RM, STRASSER M, MOORE C, et al. (2023) Evaluative criteria in practice: Findings from an analysis of evaluations published in Evaluation and Program Planning. *Evaluation and Program Planning* 97: 102226.

TONG A, SAINSBURY P and CRAIG J (2007) Consolidated criteria for reporting qualitative research (COREQ): A 32-item checklist for interviews and focus groups. *International Journal for Quality in Health Care* 19(6): 349–357.

United Nations Evaluation Group (2016) *Norms and Standards for Evaluation.* New York: UNEG.

VERHAGE A and BOELS D (2017) Critical appraisal of mixed methods research studies in a systematic scoping review on plural policing: Assessing the impact of excluding inadequately reported studies by means of a sensitivity analysis. *Quality & Quantity* 51(4): 1449–1468.

WU S, WYANT DC and FRASER MW (2016) Author guidelines for manuscripts reporting on qualitative research. *Journal of the Society for Social Work and Research* 7(2): 405–425.

YARBROUGH DB, SHULHA LM, HOPSON RK, et al. (2011) *The Program Evaluation Standards: A Guide for Evaluators and Evaluation Users*. 3rd edition. Thousand Oaks, CA: SAGE Publications.

Appendices

Appendix 1. Table Five Major Journals.

	Number	Percentage
American Journal of Evaluation	163	15
Evaluation	140	13
Evaluation Review	85	8
Evaluation and Program Planning	552	52
The Canadian Journal of Program Evaluation	130	12
Total	1070	100

Appendix 2. Full Codebook.

Domain	**Standard**	**Category**	**Source**
	Evaluand type	Program evaluation	
		Project evaluation	
		Research on evaluation	
		Evaluation theory	
		Meta evaluation	
	Sector	Development	
		Education	
		Social & Labor	
		Economic	
		Health Care	
		Organization/Administration	
		Business	
		Infrastructure	
		NGO/Participation	
		Crime	
		Other	
Investigation design and method	Evaluation purpose	Exploration	CHESS (adapted)
		Causal	
		Descriptive	
		Interpretative	

	Evaluation approach	Stakeholder Analysis Community Realist Evaluation Contribution Analysis Participatory Developmental evaluation Process Tracing Quantitative evaluation Cost-Benefit-Analysis Theory of change Formal Theory Network Analysis Experiment	CHESS (adapted)
	Meta-theory	Rationalist Constructivist Interpretative Other	
	Research design	Case Study Longitudinal Cross-sectional PTCS Mixed methods Comparison Meta-analysis Other	CHESS (adapted)
	Data collection instruments	Survey Interviews Focus groups Documents/archives Observation Other interactive/ participatory method	CHESS (adapted)
People	Affiliation	Academic Professional (independent) Business company	CHESS (adapted)

		Government Mixed	
	Evaluators' role	Author, evaluator and policymaker identical Author is evaluator, but not involved in intervention Author is not evaluator nor policymaker	CHESS (adapted)
Quality criteria	Reflexive stance	Background of author is discussed Background is not discussed	
	Empirical and methodological limitations	Discussed Not discussed	CHESS
	External validity/ generalizability	Mentioned Not mentioned	
	Selection of observations	Discussed Not discussed	CHESS (adapted)
	Choice of primary empirical source	Explained Not explained	
	Data transparency/explicitness	Shown Not shown	

Appendix 3. Table Inter-Coder Reliability.

	Percent Agreement	Krippendorff's Alpha	N Agreement vs. N Disagreement	N Cases vs. N Decisions
First Run (Deductive)	52.2%	0.453	24 vs. 22	46 vs. 92
Second Run (Deliberative)	96.4%	0.95	1059 vs. 39	1098 vs. 2196

Appendix 4. Table Co-Occurrence Table of Evaluation Categories for 1070 Articles.

	Stakeholder		**Community**		**Realist**		**Participatory**		**Developmental**	
	0	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**0**	**1**
Stakeholder										
0	937	0	771	166	906	31	900	37	911	26
1	0	133	96	37	126	7	114	19	129	4
Community										
0	771	96	867	0	834	33	831	36	844	23
1	166	37	0	203	198	5	183	20	196	7
Realist										
0	906	126	834	198	1.032	0	976	56	1.003	29
1	31	7	33	5	0	38	38	0	37	1
Participatory										
0	900	114	831	183	976	38	1.014	0	986	28
1	37	19	36	20	56	0	0	56	54	2
Developmental										
0	911	129	844	196	1.003	37	986	54	1.040	0
1	26	4	23	7	29	1	28	2	0	30

Appendix 5. Table Research Designs.

	1st codes		All codes	
	Frequency	Percent	Frequency	Percent
Case study	33	67	34	57
Longitudinal	2	4	4	7
Comparison	5	10	3	10
Mixed methods	4	8	8	13
Meta-analysis	4	8	5	8
Other	1	2	1	3
Total	49	100	60	100

Appendix 6. Table Main Type.

	1st code		All codes	
	Frequency	Percent	Frequency	Percent
Program evaluation	29	60	29	47
Project evaluation	11	23	12	19
Research on evaluation	4	8	16	26
Meta-evaluation	4	8	5	8
Total	48	100	62	100

Appendix 7. Table Metatheory.

	1st codes		All codes	
Metatheory	Frequency	Percent	Frequency	Percent
Rationalist	41	82	44	73
Constructivist	3	6	6	10
Interpretative	4	8	8	13
Other	2	4	2	3
Total	50	100	60	100

Appendix 8. Table Major Aim.

	1st codes		All codes	
	Frequency	Percent	Frequency	Percent
Exploration	19	38	22	28
Causality	17	34	28	35
Description	12	24	24	30
Interpretation	2	4	5	6
Total	50	100	79	100

Appendix 9. Articles in the Small Sample.

Adams AE, Nnawulezi NA and Vandenberg L (2015) “Expectations to Ehange” (E2C): A participatory method for facilitating stakeholder engagement with evaluation findings. *American Journal of Evaluation* 36(2): 243–255.

Anderson LA and Slonim A (2017) Perspectives on the strategic uses of concept mapping to address public health challenges. *Evaluation and Program Planning* 60: 194–201.

Bamanyaki PA and Holvoet N (2016) Integrating theory-based evaluation and process tracing in the evaluation of civil society gender budget initiatives. *Evaluation* 22(1): 72–90.

Campbell R, Townsend SM, Shaw J, et al. (2015) Can a workbook work? Examining whether a practitioner evaluation toolkit can promote instrumental use. *Evaluation and Program Planning* 52: 107–117.

Caron V, Bérubé A and Paquet A (2017) Implementation evaluation of early intensive behavioral intervention programs for children with autism spectrum disorders: A systematic review of studies in the last decade. *Evaluation and Program Planning* 62: 1–8.

Chen KH-J (2017) Contextual influence on evaluation capacity building in a rapidly changing environment under new governmental policies. *Evaluation and Program Planning* 65: 1–11.

Copestake J, Allan C, Bekkum WV, et al. (2018) Managing relationships in qualitative impact evaluation of international development: QuIP choreography as a case study. *Evaluation* 24(2): 169–184.

Crooks CV, Exner-Cortens D, Siebold W, et al. (2018) The role of relationships in collaborative partnership success: Lessons from the Alaska Fourth R project. *Evaluation and Program Planning* 67: 97–104.

Dalkin S, Lhussier M, Williams L, et al. (2018) Exploring the use of Soft Systems Methodology with realist approaches: A novel way to map programme complexity and develop and refine programme theory. *Evaluation* 24(1): 84–97.

David P and Schiff M (2015) Learning from bottom-up dissemination: Importing an evidence-based trauma intervention for infants and young children to Israel. *Evaluation and Program Planning* 53: 18–24.

Downes A, Novicki E and Howard J (2019) Using the contribution analysis approach to evaluate science impact: A case study of the National Institute for Occupational Safety and Health. *American Journal of Evaluation* 40(2): 177–189.

Frye V, Paige MQ, Gordon S, et al. (2017) Developing a community-level anti-HIV/AIDS stigma and homophobia intervention in New York City: The project CHHANGE model. *Evaluation and Program Planning* 63: 45–53.

Gosselin J, Valiquette-Tessier S-C, Vandette M-P, et al. (2015) Evaluation of a youth agency's supervision practices: A mixed-method approach. *Evaluation and Program Planning* 52: 50–60.

Ha K-M (2019) Integrating the resources of Korean disaster management research via the Johari window. *Evaluation and Program Planning* 77: 101724.

Haarich SN (2018) Building a new tool to evaluate networks and multi-stakeholder governance systems. *Evaluation* 24(2): 202–219.

Harper LM and Dickson R (2019) Using developmental evaluation principles to build capacity for knowledge mobilisation in health and social care. *Evaluation* 25(3): 330–348.

Harris K, Henderson S and Wink B (2019) Mobilising Q methodology within a realist evaluation: Lessons from an empirical study. *Evaluation* 25(4): 430–448.

Janssens FJG and Ehren MCM (2016) Toward a model of school inspections in a polycentric system. *Evaluation and Program Planning* 56: 88–98.

Jiménez-Herranz B, Manrique-Arribas JC, López-Pastor VM, et al. (2016) Transforming a municipal school sports programme through a critical communicative methodology: The role of the of advisory committee. *Evaluation and Program Planning* 58: 106–115.

Jones M, Verity F, Warin M, et al. (2016) OPALesence: Epistemological pluralism in the evaluation of a systems-wide childhood obesity prevention program. *Evaluation* 22(1). Sage Publications Sage UK: London, England: 29–48.

Kokko S and Lagerkvist CJ (2017) Using Zaltman metaphor elicitation technique to map beneficiaries' experiences and values: A case example from the sanitation sector. *American Journal of Evaluation* 38(2): 205–225.

Koleros A, Jupp D, Kirwan S, et al. (2016) Methodological considerations in evaluating long-term systems change: A case study from eastern Nepal. *American Journal of Evaluation* 37(3): 364–380.

Koper CS, Lum C and Hibdon J (2015) The uses and impacts of mobile computing technology in hot spots policing. *Evaluation Review* 39(6): 587–624.

Lawrence RB, Rallis SF, Davis LC, et al. (2018) Developmental evaluation: Bridging the gaps between proposal, program, and practice. *Evaluation* 24(1): 69–83.

Leenstra M (2018) The human factor in development cooperation: An effective way to deal with unintended effects. *Evaluation and Program Planning* 68: 218–224.

Lennie J, Tacchi J, Wilmore M, et al. (2015) A holistic, learning-centred approach to building evaluation capacity in development organizations. *Evaluation* 21(3): 325–343.

Martinaitis Ž, Christenko A and Kraučiūnienė L (2019) Evaluation systems: How do they frame, generate and use evidence? *Evaluation* 25(1): 46–61.

McIsaac J-LD, Mumtaz Z, Veugelers PJ, et al. (2015) Providing context to the implementation of health promoting schools: A case study. *Evaluation and Program Planning* 53: 65–71.

Millett LS, Ben-David V, Jonson-Reid M, et al. (2016) Understanding change among multi-problem families: Learnings from a formative program assessment. *Evaluation and Program Planning* 58: 176–183.

Milley P, Szijarto B, Svensson K, et al. (2018) The evaluation of social innovation: A review and integration of the current empirical knowledge base. *Evaluation* 24(2): 237–258.

Mohammad T, Azman A and Anderstone B (2019) The global three: A Malaysian lens on the challenges and opportunities facing restorative justice planning and implementation. *Evaluation and Program Planning* 72: 1–7.

Morgan NR, Davis KD, Richardson C, et al. (2018) Common components analysis: An adapted approach for evaluating programs. *Evaluation and Program Planning* 67: 1–9.

Najafizada SAM, Labonté R and Bourgeault IL (2017) Stakeholder's perspective: Sustainability of a community health worker program in Afghanistan. *Evaluation and Program Planning* 60: 123–129.

Nielsen JV, Bredahl TVG, Bugge A, et al. (2019) Implementation of a successful long-term school based physical education intervention: Exploring provider and programme characteristics. *Evaluation and Program Planning* 76: 101674.

Nishimura ST, Hishinuma ES, Goebert DA, et al. (2018) A model for evaluating academic research centers: Case study of the Asian/Pacific Islander Youth Violence Prevention Center. *Evaluation and Program Planning* 66: 174–182.

Noordegraaf M, Douglas S, Bos A, et al. (2017) How to evaluate the governance of transboundary problems? Assessing a national counterterrorism strategy. *Evaluation* 23(4): 389–406.

Norton S, Milat A, Edwards B, et al. (2016) Narrative review of strategies by organizations for building evaluation capacity. *Evaluation and Program Planning* 58: 1–19.

Paradis C (2016) Canada's National Alcohol Strategy: It's time to assess progress. *Canadian Journal of Program Evaluation* 31(2): 232–241.

Pouw N, Dietz T, Belemvire A, et al. (2017) Participatory assessment of development interventions: Lessons learned from a new evaluation methodology in Ghana and Burkina Faso. *American Journal of Evaluation* 38(1): 47–59.

Reeve C, Humphreys J and Wakerman J (2015) A comprehensive health service evaluation and monitoring framework. *Evaluation and Program Planning* 53: 91–98.

Richard L, Fortin-Pellerin L, Chiocchio F, et al. (2016) Création de connaissances organisationnelles à la suite d'une intervention de développement professionnel en Centre de santé et de services sociaux (CSSS) : une évaluation des laboratoires de promotion de la santé. *Canadian Journal of Program Evaluation* 31(2): 184–210.

Rolfe S (2019) Combining theories of change and realist evaluation in practice: Lessons from a research on evaluation study. *Evaluation* 25(3): 294–316.

Shmueli DF, Ben Gal M, Segal E, et al. (2019) How can regulatory systems be assessed? The case of earthquake preparedness in Israel. *Evaluation* 25(1): 80–98.

Siebert P and Myles P (2019) Eliciting and reconstructing programme theory: An exercise in translating theory into practice. *Evaluation* 25(4): 469–476.

Sokol R, Moracco B, Nelson S, et al. (2017) How local health departments work towards health equity. *Evaluation and Program Planning* 65: 117–123.

Soura BD, Bastien R and Fallu J-S (2016) Étude d'évaluabilité d'une intervention visant à prévenir l'usage de substances psychoactives lors de la transition primaire-secondaire. *Canadian Journal of Program Evaluation* 31(2): 211–231.

Sturges KM (2015) Complicity revisited: Balancing stakeholder input and roles in evaluation use. *American Journal of Evaluation* 36(4): 461–469.

Suiter SV (2017) Community health needs assessment and action planning in seven Dominican bateyes. *Evaluation and Program Planning* 60: 103–111.

Visser M, Thurman TR, Spyrelis A, et al. (2018) Development and formative evaluation of a family-centred adolescent HIV prevention programme in South Africa. *Evaluation and Program Planning* 68: 124–134.

Le LIEPP (Laboratoire interdisciplinaire d'évaluation des politiques publiques) est un laboratoire d'excellence (Labex) distingué par le jury scientifique international désigné par l'Agence nationale de la recherche (ANR). Il est financé dans le cadre du plan d'investissement France 2030 à travers l'IdEx Université Paris Cité (ANR-18-IDEX-0001).

www.sciencespo.fr/liepp

A propos de la publication

Procédure de soumission :
Rédigé par un ou plusieurs chercheurs sur un projet en cours, le *Working paper* vise à susciter la discussion scientifique et à faire progresser la connaissance sur le sujet étudié. Il est destiné à être publié dans des revues à comité de lecture (peer review) et à ce titre répond aux exigences académiques. Les textes proposés peuvent être en français ou en anglais. En début de texte doivent figurer : les auteurs et leur affiliation institutionnelle, un résumé et des mots clefs.

Le manuscrit sera adressé à : liepp@sciencespo.fr

Les opinions exprimées dans les articles ou reproduites dans les analyses n'engagent que leurs auteurs.

Sciences Po - LIEPP
27 rue Saint Guillaume
75007 Paris - France
+33(0)1.45.49.83.61
liepp@sciencespo.fr