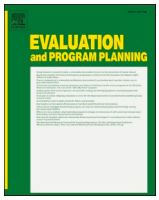
Evaluation and Program Planning 108 (2025) 102512

Contents lists available at ScienceDirect

Evaluation and Program Planning

journal homepage: www.elsevier.com/locate/evalprogplan



What evaluation criteria are used in policy evaluation research: A cross-field literature review

Céline Mavrot [a,*], Oto Potluka [b], Lars Balzer [c], Véronique Eicher [d], Sigrid Haunberger [e], Christine Heuer [f,1], François-Xavier Viallon [g]

[a] Institute of Social Sciences, University of Lausanne, Switzerland
[b] Center for Philanthropy Studies, University of Basel, Switzerland
[c] Evaluation Unit, Swiss Federal University for Vocational Education and Training, Switzerland
[d] Research and Development Unit, Swiss Federal University for Vocational Education and Training, Switzerland
[e] Institute of Management and Social Policy, Zurich University of Applied Sciences, Switzerland
[f] Evaluation and Research Service, Swiss Federal Office of Public Health, Switzerland
[g] University of Geneva, Switzerland

ARTICLE INFO

Keywords:
Policy evaluation
Evaluation criteria
Literature review
Policy fields
Evaluation theory

ABSTRACT

This literature review offers a comprehensive overview of the use of evaluation criteria across five policy fields: social services, land-use planning, teaching in higher education, vocational education, and the environment. Though it is a key part of the evaluation process, the question of how criteria are defined, chosen, and applied generates surprisingly little debate among the evaluation community. In evaluation practice, criteria are often taken for granted – and occasionally even used in ways that are neither explicit nor transparent. This cross-field literature review shows a strong presence of routinized evaluation criteria (relating to the specifics of each policy field), while some new sets of higher-degree criteria also emerge in the face of social challenges relating to sustainability, public acceptance, or social justice. Criteria development draws on both inductive bottom-up processes (which can include policy stakeholders) and top-down deductive processes (which derive criteria from the literature, as well as from national and international standards). A more profound reflection on evaluation criteria (that is, the dimensions used by societies to assess the success of policy interventions) might be required in the future of evaluation research and planning; a deeper cross-field dialogue could support this endeavor.

1. Introduction

In evaluations, criteria play an essential role in the assessment of public policies. Evaluation criteria help evaluators and planners decide whether an intervention has done its job, and whether it was successful. The standard set of criteria applied most often is either that of the Organization for Economic Cooperation and Development (OECD) and the Development Assistance Committee (DAC) (OECD, 1991), or the updated version (OECD/DAC, 2021). This comprises the widely accepted and applied general evaluation criteria long used by the evaluation community (for example, in development assistance or in evaluations of the EU Cohesion Policy): relevance, coherence, effectiveness, efficiency, impact, and sustainability. Although generally accepted and popular among evaluation commissioners, both as a complete set or as individual criteria, this set also raises several questions about evaluation criteria that are currently being discussed by the evaluation communities (e.g., at the European Evaluation Society conference).

Criticism is directed, for example, at the use of standardized evaluation criteria that consider neither the specificities of particular interventions nor the stakeholders affected by the evaluation. This controversy relates to questions of how evaluators respond to the context of the intervention and the specifics of their field, or whether evaluators apply them in the same way across evaluation fields.

Policies and their interventions affect a wide range of stakeholders

* Correspondence to: STS-Lab - Bureau, Institut des Sciences Sociales, Bâtiment Géopolis, Université de Lausanne, Quartier UNIL-Mouline, Lausanne, Vaud 1015, Switzerland.
E-mail addresses: celine.mavrot@unil.ch (C. Mavrot), oto.potluka@unibas.ch (O. Potluka), lars.balzer@sfuvet.swiss (L. Balzer), veronique.eicher@sfuvet.swiss (V. Eicher), sigrid.haunberger@zhaw.ch (S. Haunberger), christine.heuer@bag.admin.ch (C. Heuer), francois.viallon@unige.ch (F.-X. Viallon).
[1] Retired

https://doi.org/10.1016/j.evalprogplan.2024.102512
Received 3 October 2022; Received in revised form 4 October 2024; Accepted 6 October 2024
Available online 10 October 2024

who, using criteria they define for themselves, subjectively assess the usefulness of interventions. However, policies also affect whole groups of individuals collectively. How are such evaluation criteria defined? Is the definition of evaluation criteria carefully planned prior to the investigation? Are evaluation criteria explicitly mentioned, and transparently used, in the course of the investigation? In a recent review on the use of evaluation criteria in Swiss evaluation reports commissioned by federal authorities, Heuer (2017) found that that it was in only around a third of the reports examined that evaluators performed an overall assessment based on all disclosed evaluation criteria. At the very least, this shows a contrasted reality in the effective use of criteria in evaluation practice. These are questions of crucial importance both to the reliability of evaluation results and to their acceptance by stakeholders. Beyond that, we ask how field-specific criteria are used. Is there a reflexive use of evaluation criteria in the various fields? Are there some commonalities across fields? Is the use of evaluation criteria marked by innovation, or routine? Finding answers to these questions might help further develop the utility of evaluations.

To that end, we collected data from evaluations in five thematic areas and analyzed the use of criteria both within and across the specific fields. To generate results in an international context, we have focused on five thematic areas we consider relevant to current global socio-economic problems, namely: social services, land-use planning, teaching in higher education, vocational education, and the environment.

2. Evaluation criteria in evaluations

2.1. The meaning and use of evaluation criteria

Providing the grounds of an evaluative judgement, evaluation criteria set evaluation apart from other processes such as monitoring or audit (Dickinson & Adams, 2017). Many different definitions of the term "evaluation" exist. Stockmann & Meyer (2010) have identified six defining characteristics, among which choice of criteria is central. Evaluations should be carried out based on criteria explicitly related to the object to be evaluated.[1] These criteria need to be precisely defined and transparently disclosed (e.g., Widmer & De Rocchi, 2012).

The application of evaluation criteria depends on the evaluation approach used. Authors citing evaluation criteria as the basis of evaluation processes include Campbell (1991), Scriven (2015), House & Howe (1999), Weiss (1998), Stake (2004), Stufflebeam & Zhang (2017), Rossi et al. (2019), and Patton (2021). Though the terms "criteria" and "standards" are used, they have different meanings (Stake, 2004). Most evaluators use these terms for important descriptors or attributes, to help them frame the amount of that attribute needed for a certain judgment (Stake, 2004). The diversity of vocabulary in use (both within and across study fields) is a general challenge facing evaluation research (Balzer et al., 1999).

Authors rarely define precisely what they mean by an evaluation criterion. In his *Evaluation Thesaurus*, Scriven (1991) states: "In the language of evaluation, the term is… used in… a… way, to include indicators of success or merit, variables that are not part of success itself…, but rather tied to it by empirical research". Scriven (2015) states in his most recent *Key Evaluation Checklist* (KEC): "Evaluation is taken to refer to the process of determining (or the expression of a conclusion about) the goodness and/or badness, wrongness and/or rightness of something; more specifically, about the merit, worth, or significance … "an evaluation" is taken to refer to a declaration of value (…)". Based on this understanding, Scriven has approached the term "evaluation criterion" as follows: "(…) criteria of merit stem from descriptors of the evaluand – but not just any descriptor. Criteria of merit are the subset of descriptors 'that are merit-connected' (Scriven, 1980, p. 49)" (see, Shadish et al., 1991, p. 85). In 2015 (p. 4), he stated that "'dimensions of merit" (a.k.a., "criteria of merit") are the characteristics of the evaluand (X) that bear on its merit/worth/significance by definition (i.e., would typically be used in explaining what 'good X' means), and "indicators of merit" (the status of many characteristics is borderline between criteria and indicators) refers to factors that are empirically but not definitionally linked to the evaluand's merit by definition, i.e., correlates of merit."

Stake and Davidson define "evaluation criteria" thus: "a criterion is an attribute of an object or activity used to acknowledge its merit and shortcoming. It can be a trait or ingredient seen to be essential. It becomes a basis for judgment or action when a standard is set." (Stake, 2004). Davidson (2005a) understands the evaluative criteria or dimensions of merit as "attributes (e.g., features, impacts) of the evaluand that we look at to see how good (or how valuable, how effective, etc.) it is." In the *Encyclopedia of Evaluation*, she summarizes her understanding of the term as follows: "The aspects, qualities, or dimensions that distinguish a more meritorious or valuable evaluand (…)" (Davidson, 2005b).

For Stufflebeam (2001), criteria are "standards on which to base judgments." To him, "values" are "principles, attributes, or qualities held to be intrinsically good, desirable, important, and of general worth." Statements by Weiss (1998) can also be interpreted in this way; she writes about "standards of judgment" in connection with criteria. In 2017, Stufflebeam & Zhang (2017) define criteria as "explicit variables and interpretation rules, for use in assessing and judging a program".

These approaches to defining evaluation criteria thus open up the following definitional field. "Evaluation criteria" can be synonymous with: dimensions of merit; a subset of merit-connected descriptor; characteristics of the evaluand (X) having a bearing on its merit/worth/significance; attributes of an object or activity used to acknowledge its merits and shortcomings; attributes (e.g., features, impacts) of the evaluand that we look at to see how good (or valuable, effective, etc.) it is, or standards on which to base judgments and explicit variables and interpretation rules, to be used in assessment procedures (see also Alkin, 2013).

2.2. Developing evaluation criteria

The importance and purpose of evaluation criteria might be clear to evaluators; this may also be why they do not always explicitly define them. However, how evaluators understand and define them often remains unclear. There is widespread agreement in the literature that evaluation criteria should be established prior to the start of an evaluation (Alkin et al., 2012; Scriven, 2007; Stake, 2004), though the ways of doing so differ. There is a danger that "the conventional evaluation question entails the domination of certain criteria and values and consequently the exclusion of other criteria, values, voices and people" (Abma, 2000: 199). Scriven (2007) believes that evaluators should collect the values relevant to the object of evaluation and derive evaluation criteria from them, while Stufflebeam (2001) involves stakeholders directly. According to Stake (2004), evaluators "do ask staff people and other stakeholders, sometimes participants and recipients, for help in identifying criteria." However, he acknowledges that "sometimes the evaluators are in a position to know better than anyone else the relevant criteria, but often others nearby, with acuity and legitimacy, are better at clarifying the standards in use." Both Scriven (2007) and Stufflebeam (2001) have developed checklists to help evaluators identify the values of the evaluation object and evaluation criteria, and update these on an ongoing basis.

Davidson (2005a) analyzes existing proposals on how to identify evaluation criteria and compiles basic concepts and tools that are either essential:

- A needs assessment (identification of those impacted and their needs, which are synonymous with the outcome criteria; definition of both context and performance needs)
- A simple logic model that links the evaluand to the needs

[1] In this article, we use the verbs to *assess* and to *evaluate* as synonyms.

- An assessment of criteria for other relevant values
- A checklist for thinking of other relevant criteria, using the headings of Process, Outcome, and Costs (adapted from Scriven)
- A strategy for organizing the criterion checklist

Back in 1978, Rycroft identified possible sources for defining evaluation criteria for a given study, which can be mixed and include: scientific literature, policy field evidence (drawn from the past or from international data), participation of policy stakeholders in general, and policy targets. However, alongside these theoretical developments in the evaluation literature, there is a gap in the ways criteria are actually used in evaluation practice. Below, we review the evaluation literature in five policy fields through the lenses of definition, choice, and operationalization of criteria.

3. Research questions: what evaluation criteria (if any) are used – and how are they used?

In this literature review, the project team answered the following overarching questions in the selected policy fields.

Definition. How do authors understand and define evaluation criteria?

- *Development*: What process do authors follow to determine the criteria?
- *Content*: Which concrete evaluation criteria are used?
- *Application*: How are the evaluation criteria used (as indicators, with threshold values, etc.)?

The adopted theoretical approach is abductive (Tavory & Timmermans, 2014), with an iterative confrontation between general deductive expectations as to what the use of evaluation criteria is expected to look like based on evaluation literature and the effective use of criteria among policy fields observed through the inductive literature review. Our theory-based expectations are about the explicit mention of criteria, the existence of a definition and the possible ways to construct and apply them; the expectations are not about the content of criteria, that can vary widely depending on the field.

4. Methods, data, analysis

4.1. Literature review

4.1.1. General framework

With the number of scientific publications available increasing rapidly each year, it is useful to summarize individual studies on the same topic in a concise manner, in order to provide an overview. In line with literature review standards, we adopted a systematic approach to the selection, inclusion and analysis of the literature sources, including the definition of a study protocol as well as 'a priori' defined literature databases and inclusion criteria (Machi & McEvoy, 2016; Ressing et al., 2009). Ultimately, the results were summarized in a qualitative manner. During the period between June 2020 and November 2020, various scientific literature databases were searched (such as Scopus, Web of Science, SocINDEX, SSOAR), using specific search combinations concerning evaluation criteria) in five different policy fields (for more details, see Appendix A). The research team is made of experts respectively specialized in each of the analyzed policy fields (vocational education, social services, land-use planning, etc.). A preliminary scoping of the literature showed that the use of the same databases across policy fields was not suitable to identify the articles that related to the research question (i.e., what use of evaluation criteria in evaluations studies within the respective fields). Relying on the field-specific expertise of the research team members, the most relevant databases have therefore been selected inductively for each policy field, by identifying which outlet focused on the reporting of evaluation studies (and from that, in which database these outlets were indexed). The aim was to strike a balance between systematicity and adequacy, without losing the requirements of context-sensitive analysis.

The same explanation holds for the choice of the keywords used for the literature review in each policy field. While the core keyword strategy was to search for "evaluation criteria" AND "name(s) of the policy field", adaptations were made for one policy field after the preliminary search steps of the review. Indeed, the results obtained from the approach with the two search terms "evaluation criteria" and "land-use planning" were highly heterogeneous and with no clear connection to policy evaluation in the field of land-use planning. Therefore, a third criterion was added: "policy evaluation". The reason is that historically, several other forms of studies are performed in the land-use planning field, such as GIS-based analyses and impact assessments, which were often dissociated from a policy or its political context, and thus were less directly related to our research question about policy evaluation. With a too wide keyword research strategy, the results would have therefore been poorly comparable with the other policy fields. The search initially resulted in a total of 547 hits, of which 179 literature sources remained as the basis for analysis (after excluding duplications and considering the inclusion criteria). The further structure of the contribution follows the steps of a literature review proposed by Machi & McEvoy (2016).

4.1.2. Literature search

The five policy fields included in this literature review are: social services, land-use planning, teaching in higher education, vocational education, and the environment; these were selected for their central relevance to policy evaluation research as fields in which evaluation research is particularly prolific. Moreover, between them, these fields cover a substantial proportion of the policy research area. We have selected the investigated themes according to three dimensions: i) their societal relevance; ii) their relevance to policy evaluation theory and praxis; and iii) their suitability regarding our research question. Regarding societal relevance, although there are differences in expenditure priorities across countries, the OECD mentions social protection, health care, and education as, on average, the most important spending categories in OECD countries (OECD, 2021, p. 84). Moreover, combating climate change through environmental protection (including land-use planning) belongs to the Green Deal, a flagship of EU policy, underlining its importance (EC, 2019). Regarding relevance to evaluation, the five chosen policy fields have a long-standing evaluation tradition. This makes them comparable in analyzing the degree of sophistication achieved regarding the use of evaluation criteria in these fields. While we acknowledge that other fields like health or development cooperation have a high societal and evaluation relevance, we excluded them because of considerations regarding suitability to the research question. We hold that the inclusion of these fields would have induced a bias in the results, because they have achieved a high field-specific standardization regarding evaluation criteria. In the field of health, evaluation criteria are strongly consolidated around, for instance, the question of efficacy and safety regarding drug authorizations or health interventions, among others through randomized controlled trials, or health outcomes measured through mortality rates. As to development cooperation and assistance, the standardization stems from international standards regulating this field, as opposed to the selected policy fields that strongly remain within the realm of national governance.

Since each policy field has its own publishing bodies, we have selected relevant bibliographic databases that refer specifically to the respective fields. Further literature was also taken into account via searches in Google Scholar. The policy fields were studied using a uniform search strategy: inclusion criteria for the articles in the literature review, and analysis of each field along the research questions (abovementionned, Section 3). The searched databases for each policy field are presented in Appendix A. The literature review includes article published until November 2020. The inclusion criteria are presented in Appendix B.

4.2. Article selection process

Fig. 1 shows the literature review selection tree for all policy fields. The sources initially found were reduced by the inclusion criteria in every policy field. A total of 179 literature sources remained, divided between the policy fields as follows: 34 social services, 52 land-use planning, 22 teaching in higher education, 19 vocational education, and 47 environment.

Design according to Lacouture et al. (2015)

Appendix C contains a detailed presentation of the characteristics of the analyzed articles for each policy field.

5. Results: the use of evaluation criteria across policy fields

5.1. Social services

More than three-quarters of the 34 articles fail to explicitly define "evaluation criteria" in this field. Some authors complain that no consensus has been reached on the question (Min & Huilan, 2020). Some fairly general attempts to define evaluation criteria can be found, for example in the sense of a defined standard (Kagle, 1979), a rating of existing health and social indicators (Daniel et al., 2009) or in terms of perceived outcomes (Sørensen & Bay, 2002). More often, reasons are given as to why criteria are needed, or what could be done with them. Evaluation criteria "can productively inform policy and practice when actors deliberate on how to assess and improve" the evaluation object (Hanberger et al., 2016, p. 675), or "reflect the roles and interests of the stakeholders in question" (Thomas & Palfrey, 1996). It is only in a few exceptional cases that definitions are provided with reference to authors, standard literature or evaluation models – for example, the Kirkpatrick Evaluation Model (Pratama & Setiawan, 2018), or Ball's (1988) standard *Evaluation in the Voluntary Sector*.

About a quarter of all evaluation criteria used in these contributions were developed by the authors. Criteria are thus often developed in an iterative process involving various stakeholders – very often in focus groups (Chipman et al., 2002; Hanberger et al., 2016; Kagle, 1979; Moro et al., 2007). The methodological approach often includes a multi-stage survey procedure, with feedback loops being used to develop and refine evaluation criteria to reach consensus (Chipman et al., 2002; Gibney et al., 2019; Kastein et al., 1993). Furthermore, by starting from stakeholder problems and experiences (Moro et al., 2007) and by not identifying their evaluation criteria within the context of any theoretical analysis or benchmarking, they follow the responsive evaluation perspective put forward by Stake (1995). Another article refers to stakeholder-focused criteria (Thomas & Palfrey, 1996) and identifies three relevant stakeholder groups: funders (the state, insurers, etc.), service users, and service providers (professionals, volunteers, and managers), whose resources can be pooled to define which criteria are significant (Langer et al., 2019). We observe, then, that a participative approach to criteria development is significant in this policy field. A further quarter of the contributions refer to evaluation criteria without giving information on the background of their development or origin.

Some authors embed their evaluation criteria adjacent to theories and/or evaluation models. A few authors derive their evaluation criteria from existing policy standards/overall objectives – for example, from the National Association of Social Workers' *Standards for Social Work* (Kagle, 1979), the U.S. State Department of Education (Garey, 2002), the *Hospital Accreditation and Evaluation Program* (Song et al., 2019), or the OECD's DAC (Hideg, 2019). Other authors selected their evaluation criteria on the basis of systematic literature reviews.

In about three-quarters of contributions, no target values or threshold values for evaluation criteria are given. In those contributions in which thresholds are discussed, this is done quantitatively – suggesting percentages, statistical significance tests between groups, (self-assessment) scales and scores. In a quarter of the contributions, no concrete evaluation criteria were used at all. The choice of criteria tends to be project related. Criteria relating to behavioral change (Gilgun, 1988) can be identified: change of attitudes, change of skills (Pratama & Setiawan, 2018); satisfaction, knowledge, reaction (Clarke, 2001); child safety, physical care, emotional care, and support (Chipman et al., 2002) as well as psychosocial functioning, life satisfaction, and costs (Schmidt-Posner & Jerrell, 1998). Depending on context, "criteria such

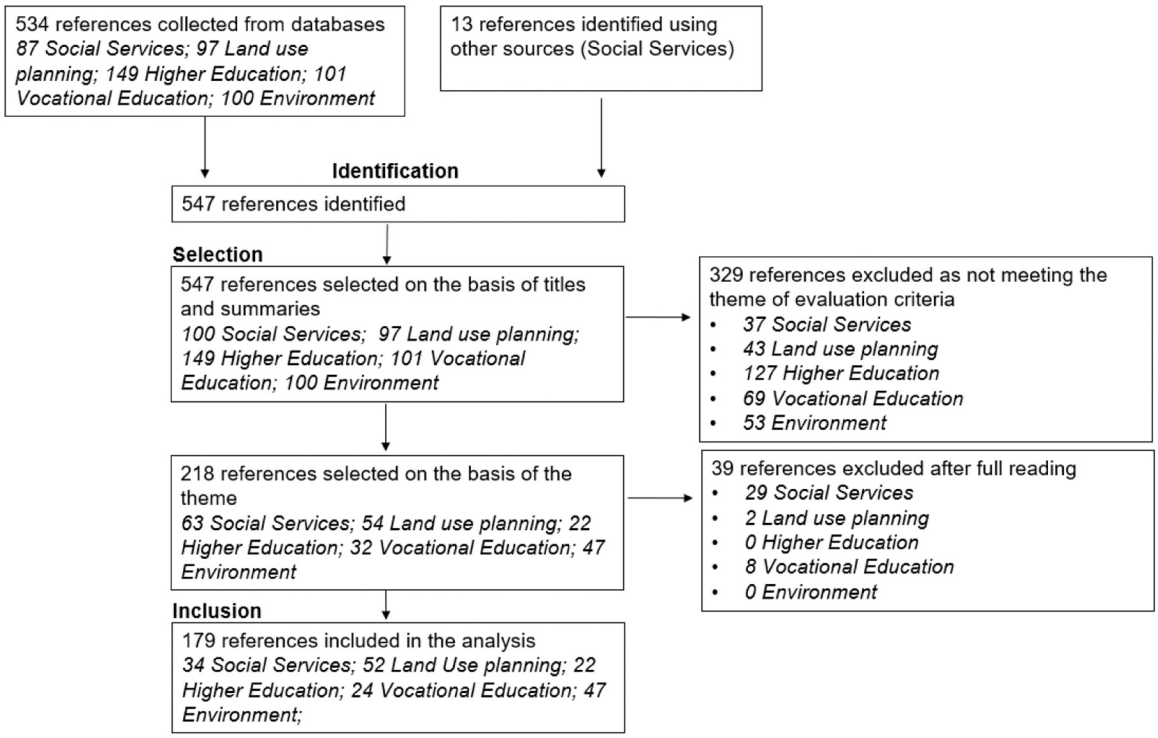


Fig. 1. Literature Review Selection Tree.

as effectiveness and appropriateness are likely to be the ones emphasized by professionals. Managers … are more likely to concentrate on efficiency and accountability" (Thomas & Palfrey, 1996, p. 128).

Some authors adopt a more comprehensive perspective, providing a classification of criteria according to different levels: "micro evaluation criteria (service transaction and service delivery): amount of volume, timelines, availability or continuity, friendliness, empathy; Meso evaluation criteria (effectiveness of the service provided): specific policy goal achieved, users' needs met; macro evaluation criteria: services are in accordance with general social values and norms, public good objectives" (Rieper & Mayne, 1998, p. 120). On the whole, others relate evaluation criteria to social justice indicators: equitable education, access to the labor market, poverty prevention, intergenerational justice, social cohesion, and non-discrimination (Zamfir, 2017). Eeckhout et al. (1996) also include the alleviation of poverty, women's emancipation, and environment protection. Similar examples can also be found in the study by Sørensen & Bay (2002, p. 375) who use (among other criteria) the assessment of equality, working conditions, and democratic governance. In this policy field, a particularly comprehensive approach is taken to evaluation criteria – which are sometimes conceptualized at a societal and democratic level.

Very few authors reflect on the theoretical development of evaluation criteria in the social services policy field, their use, and their necessary requirements. Some complain of the lack of consensus regarding appropriate alternative measures of effectiveness and call for the identification of basic parameters as well as academic support (for example, Clarke, 2001; Min & Huilan, 2020; Moro et al., 2007; Voisin & Berringer, 2015). The question of whether the criteria used are reliable, valid, and reasonable often remains unanswered (Waterhouse & Carnie, 1992), and the possibility of criteria misuse is also pointed out. Politicians "will understandably vary in their preoccupations with criteria depending on, among other things, whether they belong to the party in power at the time in question" (Thomas & Palfrey, 1996).

5.2. Land-use planning

Of 52 research articles, twelve provide a definition of evaluation criteria, which are considered as a way to "evaluate the adequacy of different alternatives" (Langemeyer et al., 2016), "assess the quality of [planning] products and processes" (Oliveira & Pinho, 2010), or "allow for systematic analysis of plan implementation" (Muñoz Gielen & Mualam, 2019).

Considering the sources of evaluation criteria definition, 34 papers refer to the literature in the policy field or in policy evaluation in general (e.g., Mazmanian & Sabatier, 1983a). Seven papers derive their criteria from policy goals (e.g., planning legislation, plan goals). Seven others refer to the researchers' own criteria, or to those of experts, developed specifically for the study. Three papers refer to criteria developed by stakeholders involved in the evaluation process, and one to criteria defined by decision analysis software.

The effective use of evaluation criteria is very high in the land-use planning policy field, with 45 studies reporting the application of evaluation criteria in their research, while the remaining articles focus on theoretical and/or methodological aspects of evaluation. Of these 45 qualitative and quantitative papers, 25 apply the criteria together (using indicators measuring specific elements of the object under study) while 20 others use evaluation criteria directly as indicators.

Land-use planning literature sets out two broad approaches to land-use policy evaluation: conformance- and performance-based. Conformance-based approaches focus on questions of effectiveness (differences between intended policy objectives and outcomes), while performance-based approaches appraise all impacts of plan implementation, including both intended and unintended effects (Shahab et al., 2019). Additional overarching criteria that may be considered are efficiency, equity, social and political acceptability, and institutional arrangements (administrative feasibility, transaction costs). Depending on the research topic considered (e.g., land (re-) development, transport planning, energy planning, land or shore protection), the range of indicators considered may vary considerably. Such variation particularly applies to multi-criteria decision analysis – a data driven methodology that may (depending on available data and goals of the study) include numerous indicators covering ecological, locational, topographic, economic, climatic, zoning- and property-related aspects. Land-use planning evaluations also include cost benefit analyses, which focus on socio-economic indicators such as employment, private and public income, or sales.

Regarding the object of evaluation, two broad types of land-use planning evaluations may be distinguished. One type covers evaluations at a strategic level, focusing on the political-administrative aspects of land-use plan implementation such as the definition of planning goals and the planning process itself. These studies can include evaluations of national plans, strategic environmental assessments, or studies of plans for metropolitan areas. The other type of evaluation covers the more operational level, including land-use plans with (more) binding constraints on landowners and consequent land-use changes. The operationalization of the evaluation criteria varies between studies, depending on the goals of the study and the object to be evaluated (e.g., stakeholder involvement, accessibility, natural hazards, protection of historical sites, sustainability).

To sum up, most papers on land-use planning evaluation rely on the literature to define their evaluation criteria, whereas theoretical papers both provide definitions of evaluation criteria and propose overarching criteria to be used in land-use planning evaluations. Methodological and empirical papers, on the other hand, often do not show a clear understanding of evaluation criteria. Such confusion is reflected in two ways: first, terminology use is not always consistent, so that criteria may also be referred to as 'dimensions,' 'indicators,' or 'best practice principles.' Second, when referring to the measurement of land-use policy outputs, impacts or outcomes, usage of the terms 'evaluation criteria' and 'indicators' often overlaps. The reasons for these inconsistencies are complex; one possible explanation could lie in the heterogeneous disciplinary backgrounds of researchers involved in land-use planning evaluation. These researchers bring different theoretical perspectives (e.g. from planning, geography, ecology, engineering, or architecture) into evaluation practice which may enrich evaluation research, though they may not be well-acquainted with the literature on public policy analysis and its terminology (e.g., Knoepfel et al., 2007; Mazmanian & Sabatier, 1983b). In other words, the definition of evaluation criteria often lacks clear theoretical foundations. More importantly, the selection of evaluation criteria frequently strays into questions of measurement of these criteria and/or data availability. Conclusions as to the meaningfulness or relevance of evaluation criteria often remain unsatisfactory, particularly in multi-criteria decision analyses. There is a tendency to consider the choice of criteria and/or indicators as primarily one of a technical nature. According to the literature on policy instrument implementation (Knoepfel et al., 2007; Linder & Peters, 1998; Salamon, 2002; Vedung, 1998), choice of criteria are subject to debate among the policy actors involved, precisely because these criteria allow a judgment on past/upcoming planning decisions to be made. The papers analyzed tend to either ignore or technicize such choices.

5.3. Teaching in higher education

The literature sources of the sample related to higher education focus on the evaluation of teaching. However, since we are exclusively interested in which criteria were used to evaluate the quality of teaching in higher education, the (many) papers focusing on the evaluation of teaching as a criterion for personnel decisions were explicitly excluded from the sample. A first result is that none of the authors of the 22 selected publications includes either a theoretical definition of evaluation criteria or a definition what constitutes an evaluation criteria. In one paper, the author does define what constitutes an indicator: "not all

data are indicators; (…). An indicator acts as a signal, an indicator, which requires an important activity of interpretation in order to reach a judgment that should never be a substitute for it" (Romainville, 1999, p. 418). In one paper (Arthur et al., 2003), the authors test the distinction between Kirkpatrick's Level I (reaction) and Level II (learning) evaluation criteria and thus explicitly refer to his model (Kirkpatrick, 1976, 1996). None of the remaining 20 papers include either any reference to evaluation theories or any definition of criteria or indicators.

Processes used to determine the evaluation criteria were diverse: about one-third of the papers did so via reference to previous research and literature reviews, while another third applied the evaluation criteria used in their own higher education institutions (some of which have been drawn from the *Individual Development and Educational Assessment* (IDEA) rating system). Three papers based the development of the evaluation criteria on a theory or model: the *Presage, Process, Product Model* (Biggs, 1993), the conceptual framework of multidimensional learners' evaluation (Sritanyarat, 2014), and Kirkpatrick's four level model (Kirkpatrick, 1976, 1996). In three additional papers, evaluation criteria are developed in working groups that include a range of sources such as students (Amin, 2002), faculty (Baker et al., 2015), or expert and novice teachers (Dunkin, 1995).

The evaluation criteria are mainly (in two-thirds of the publications) used as indicators without specified thresholds. In one study (Del Carmen Bas et al., 2017), the authors propose a composite index of evaluation criteria, in which the criteria are weighted differently based on the opinion of experts from a Teaching Evaluation Committee and underlying correlations between criteria. Just two of the 22 publications include thresholds defined for the indicators (Baker et al., 2015; Essack et al., 2012). In the study by Baker et al. (2015), students were asked to rate items on a scale from 1 to 7, with a threshold for a satisfactory rating. In the study by Essack et al. (2012), indicators with thresholds are limited to objective criteria – specifically, graduation rates in study programs and pass rates in modules. Four publications defined no specific criteria, mentioning only dimensions.

The evaluation criteria used in the publications are similar across the publications, due to the fairly homogenic nature of teaching evaluation. Specific evaluation criteria were mentioned in 18 of the 22 publications, and in all of these, the evaluation criteria included a rating of the teacher (e.g., overall rating of the teacher, rating of teacher's rapport with students, teacher's enthusiasm). A third of these publications used students' self-perceived learning, perceived workload, and/or rating of the course content and objectives as evaluation criteria. Two publications included quantitative criteria such as drop-out or pass rates.

Overall, we see that in the overwhelming majority of publications regarding the evaluation of teaching in higher education, authors neither define what they mean by evaluation criteria, nor rely on a specific evaluation theory. In most publications, however, specific evaluation criteria are named for the evaluation of teaching – and these are often similar. Teachers are evaluated, and students used as (at least one) source, in all of the publications. Criteria are used repeatedly in many of the publications, giving the impression of a fairly uniform use of criteria in the evaluation of teaching in higher education. The publications analyzed here date from 1988 to 2020, and during that period, the criteria did not change systematically (in terms of either content or source). This shows that, although authors do not explicitly state what they mean by evaluation criteria, they do rely on similar ones; this is in part due to the fact that they use those found in literature and existing instruments in higher education.

5.4. Vocational education training

Though strict criteria were used for their selection, the 24 articles found in the field of vocational education training are quite diverse. Evaluation objects cover very different fields of interest, including a range of learning settings, educational material, educational programs, and schools. With two notable exceptions, no theoretical definition of evaluation criteria is given. Only Höhns (2017, p. 327) labels them as "narrations about feedback, criticism and suggestions for improvement", while Khaleel (1988) describes them as external and internal factors that determine program quality. However, neither of these is a strict technical definition. All other texts just use evaluation criteria as taken for granted. The process used to determine evaluation criteria is largely based on literature – but this literature is mainly related to education, rather than evaluation. The sole exception is Custer et al. (1997), who adapted Daniel Stufflebeam's CIPP-model (interestingly, without citing any of the author's actual text).

Use of criteria is made as a function of indicators, covering the field of the diverse evaluation objects; consequently, no common set of criteria can be found. Almost every text uses its own set of criteria. Empirically-oriented texts adapt them to their specific situation, e.g., satisfaction with usability and utility, support (Beckers et al., 2019), fitness for purpose, cognitive complexity, self-assessment, authenticity, transparency, comparability, reproducibility of decisions, fairness, acceptability, meaningfulness, educational consequences, costs and efficiency (Baartman et al., 2013), competence and task mastery (Kollöffel & de Jong, 2016), classification status (Harth & Hemker, 2013), post-prison employment as a rehabilitation measure (Gleason, 1986), or multiple effectiveness criteria (De Maeyer et al., 2010). Theoretically-oriented texts (aiming, for instance, to develop possible lists) emphasize the broad variability of criteria and "use a flexible body of different types of criteria which are adjusted to the situation at hand" (De Vos et al., 2019, p. 702). The idea of explicitly defining criteria (preferably prior to data collection) is absent. In the empirically oriented texts, where criteria are used as a basis for judgment and valuing, they do so as an act of statistical comparison between groups, or as scale characteristics. In summary, evaluation criteria in the field of vocational education training are generally derived deductively from the existing education literature and are not subject to substantial theoretical reflections in the analyzed studies. Furthermore, a great diversity of criteria is in use, depending on the specific object under evaluation.

5.5. Environment

The studies in the evaluation of environment protection show three characteristics, the first of which concerns the vague nature of theoretical definitions of evaluation criteria. Usually, authors define the criteria as given, assuming readers are familiar with them, or refer to the literature that defines the criteria. Second, the criteria and their weights are defined in a participatory manner with local stakeholders in a minority of cases; in other instances, the authors have made these decisions themselves. Third, criteria are used in complex models, with weightings applied to them according to their relative importance. The use of criteria in multi-criteria analysis creates an aggregate indicator enabling comparison between a range of possible solutions (investment in various places, or approaches to an environmental problem). The majority of studies provide "sophisticated" information for decision-making.

A large majority of the 47 articles understand evaluation as an analytical instrument designed to measure the *impact*, *effects*, or *performance* of a policy. This indicates a fairly instrumental and straightforward understanding of the role of evaluation criteria, which are used to assess both the added value of the policy and the fulfillment of its objectives. Three studies insist on the need to use evaluation criteria to assess processes (e.g., the quality of deliberation processes in participative environmental policies, Van Den Hove, 2000). Strikingly, very few of these articles fail to explicitly mention evaluation criteria, even though, for about a third of them, the definition of evaluation criteria is mainly implicit.

In terms of the process through which the authors arrive at their evaluation criteria, the articles are split between two groups. For about half of the articles, evaluation criteria have to be defined in an ad hoc manner, depending on the specificities of both study field and program (Bellamy et al., 2001). For these authors, the process of criteria

development has to be both iterative and achieved through empirical field exploration (for instance, via interviews and documentary analysis, e.g. Holmes & Clark, 2008). For the other roughly half of the sample, evaluation criteria must build on the results of past studies, and researchers are able to rely on existing criteria taxonomies, which have the advantages of being robust and allowing sound comparisons (Barron & Ng, 1996). This relates to the traditional divide between deductive and inductive research streams.

In terms of how the evaluation criteria are applied, the set of articles shows more variety. Some authors use criteria in a highly systematized and formalized manner, defining scales, attributing scores, and modeling the results in mathematical equations (Portney & Stavins, 1994). Some other authors, however, apply the criteria in a qualitative fashion.

Regarding the kind of evaluation criteria applied in the field of environmental policy, the traditional effectiveness and efficiency criteria are used in several of the studies (Cabugueira 2011; Goulder & Parry, 2008; Gysen et al., 2006). Interestingly, discussions are taking place as to the specificities of assessing programs and policies that aim to preserve a common good – such as the environment. In the wake of these reflections, several articles propose further society-owned criteria, such as social effectiveness (Gysen et al. 2006), equity and capacity-building (Bellamy et al., 2001), public acceptance (Richards, 2000), transparency (Kunseler & Vasileiadou, 2016), community acceptability (Gunningham & Young, 1997), or social costs (Kim, 2007). Finally, because of the specificities of this policy field, some evaluation criteria are technical and relate to the nature of the investigated matter (for instance risk level, or degree of resource use, e.g., Munda et al., 1994). This can explain the (sometimes high) degree of formalization found in this area, in terms of how criteria are applied.

In this policy field, there are specific calls for future research on evaluation criteria. One article draws attention to the fact that complex trade-offs exist between various evaluation criteria – in particular, between damage caused to the environment and monetary value (Kim, 2007). Another stresses that when dealing with contentious issues such as environmental policies, best practice is to explicitly define criteria through an open process (Barron & Ng, 1996). The openness of environmental evaluation toward public value is summed up by (Kunseler & Vasileiadou, 2016) as follows: “stakeholders with diverse interests on environmental policy may help supplant the classic-rationalistic principles of effectiveness and efficiency to help develop new evaluation criteria based on good governance principles (i.e., participation, transparency and fairness)”. In short, evaluation criteria are both used and explicitly named in the vast majority of studies in this field. These criteria are developed through both inductive and deductive processes, covering the classical aspects of efficiency and effectiveness, the technical aspects of environment policies, and more innovative features relating to democratic processes and acceptance.

6. Discussion

Coming back to the research questions lying at the heart of this investigation, it is striking to note that evaluation research in the different policy fields almost fails to address the question of the theoretical definition of criteria. Evaluation criteria are mentioned as though based on a commonly shared understanding of the notion, and only rarely is a definitional effort made. More information can be found on the process through which the authors come to develop or select the criteria relevant to their investigation. Depending on the policy context, criteria selection can inductively rely on participative processes involving relevant stakeholders, or deductively draw from past research or institutional, national, or international guidelines. The degree of formalization through which criteria are analytically applied (scores, thresholds) is dependent on policy field, the existence of commonly recognized standards, and the level of technicality of the policy in question. Lastly, as to the content of the criteria, the literature review uncovers a wealth of diversity, though we have been able to highlight some superordinate categories in each field, as shown in Table 1 below.

Hence, Table 1 provides examples of commonly used evaluation criteria (i.e., found several times in the literature review) in the five analyzed policy fields. Hence, the content of Table 1 is based on the inductive results of the literature review. Given the lack of theorization and the overall heterogeneity of evaluation criteria in the literature, this overview provides a first step for a future reflection on this core evaluation issue. The Table also shows that beyond the diversity of evaluation criteria, transversal element can be found.

As shown in the Table, evaluation criteria in each field can be regrouped in overarching categories that were inductively found in the literature review, for instance: processes, democratic dimensions, behavioral-change dimension, but differ across fields of analysis. These overarching categories in the Table are constructed by us and marked in italics. They pertain to various analytical levels and might help future developments on the conceptualization of evaluation criteria. Finally, while policy evaluation in various fields is mostly conducted in silo, the table comparatively shows the criteria used across fields. This can help avoiding a tunnel vision and provide cross-field inspiration, as criteria commonly used in one field could be interesting to import in another one.

This literature review thus offers a contrasted perspective on the status of criteria in evaluation research. In most of the policy fields reviewed, evaluation criteria are actually named and applied. Yet in the field of social services, a quarter of the studies reported no criteria at all, which is surprising and goes against the requirement of transparency and systematicity laid out in evaluation research. Overall, evaluation criteria are mostly used in a transparent way, fulfilling a specific aim in the research procedure such as the usefulness of social services (welfare), strategic studies (land-use planning), teachers’ rating (higher education), assessment of learning settings (vocational education), or impact measurement (environmental policies). Interestingly, though evaluation criteria are used in most studies, few dwell on the question of their definition or status from a theoretical perspective. We should beware of the fact that there is more to evaluation criteria than simply reducing them to the "program goals", and to subsuming the latter under the concept of "effectiveness" (Weiss, 1993: 96). The problem in adopting such a perspective is that it leads the evaluator to "accept the premises underlying the program" (Weiss, 1993: 100), thus making evaluation lose its critical function.

Choices of evaluation criteria show consequent variations, both within and across policy fields. In some areas, such as social services, the focus lies on iterative processes and criteria tend to be defined inductively on a case-to-case basis – sometimes by including policy users themselves. At the other end of the spectrum, the criteria applied in land-use planning are mainly derived from the literature, though they also take into account specific policy objectives. In vocational education, the process is usually literature-based, though criteria are ultimately also adopted on an ad hoc basis. Higher education and the environment turn to different criteria development strategies, drawing on pre-existing studies, institutional standards, or existing taxonomies.

Acknowledging the democratic dimension of evaluations, the social services field, and to a lesser extent, the environment field, place special emphasis on participative processes and the inclusive development of evaluation criteria. Careful reflections are carried out regarding policy effects and legitimacy for their stakeholders. While this is undoubtedly related to the particularities of these fields (in which the production of a common good is stressed) it is intriguing to observe that inclusive processes are less present in the other fields. Certainly, there is no panacea in evaluation, and both participatory and non-participatory approaches have benefits and drawbacks. Expert top-down evaluation procedures are based on rigorousness and independence and thus enjoy results-based legitimacy, while their bottom-up counterparts valorize stakeholder-ownership, and enjoy process-based legitimacy (Sager & Mavrot, 2021). In any case, one question that may well be worthy of

Table 1
Example of commonly used evaluation criteria in each field (selection).

Social Services	Land-Use Planning	Teaching in Higher Education	Vocational Education	Environment
Behavioral change criteria	*Formal Dimensions*	*Teacher-Focused, e.g.*	*Actor-Focused*	*Material Dimensions*
Change of attitudes	Conformance, Compliance	Competence	Satisfaction	Productivity
Change of skills		Clarity	Competence / Skills	Profitability
Child safety		Enthusiasm	Performance	Employment
Physical care			Interpersonal iterations	
Emotional care & support			Prior knowledge	
Psychosocial functioning				
Life satisfaction				
Merit criteria	*Substantive Dimensions*	*Student-Focused, e.g.*	*Project-Focused*	*Democratic Dimensions*
Technical quality Objectivity	Impact on target group, Performance	Engagement	Effectiveness	Quality of life
Validity	Effectiveness: contribution to resolution of the public policy problem	Motivation	Utilization	Equity
Worth criteria		Learning gain	Goal attainment	Participation
Utility	Procedural effectiveness		Quality	Support
	Efficiency		Clarity	Fairness
			Coherence	Legality
			Long-term impact	
			Project success	
Social justice related criteria	*Sustainability dimensions*	*Course-Focused, e.g.*		*Process*
Equitable education	Distribution of economic, social, and environmental benefits Visual impacts, Impact on landscape	Content		Traceability
Access on the labor market		(Perceived) Utility		Adequacy
Poverty prevention				Capacity
Intergenerational justice				Partnership
Social cohesion				
Non-discrimination				
General criteria	*Processual dimensions*			*Effects*
Relevance	Participation, Inclusion of stakeholders			Goal attainment
Effectiveness	Legitimacy (acceptability, equity)			Improvement
Efficiency				
Implementation				
Impact				
Sustainability				
Appropriateness				
Accountability				

Note: The overarching categories are marked in italics.

more systematic reflection is that of which path should be adopted in the criteria development process, according to the specifics of each field.

A further dividing line across policy fields relates to how criteria are applied to answering evaluation questions. There is a mix of qualitative and quantitative approaches to the use of criteria, with land-use planning and environment evaluations representing the most formalized fields from this perspective. In these (often highly technical) studies, evaluations are often performed using formalized processes involving scales, scores and threshold models, or multi-criteria evaluation models with aggregate indicators (in the environmental field). In the fields of higher education and land-use planning, we find instances of the notions of "criteria" and "indicators" being used interchangeably, even though, in our understanding, these are two distinct instruments of the research apparatus. Criteria provide the gauges aimed at structuring judgment on a phenomenon, while indicators are the concrete set of data providing the basis for assessment. From this perspective, evaluation criteria pertain to the theoretical dimension of the studies, whereas indicators fall within the realm of methodology.

From a transversal perspective, the literature review also reveals that few cross-field evaluation criteria exist. There is no one-size-fits-all approach, and field specificities trump possible evaluation standards. Higher education and social services usually adopt a user-centered approach at the individual level, with a focus on self-perceived utility and satisfaction as well as change (learning, behavior change). Evaluations in the field of vocational training also assess individual parameters – such as intervention meaningfulness and rise in user competency – while aggregating the results at a higher level to formulate a judgment on dimensions such as employability or costs. Lastly, with larger scale policies, environmental and land-use studies adopt a broader focus on capacity-building or feasibility. These two fields also adopt innovative approaches that value the societal dimension of acceptability, as do social services studies, which sometimes apply criteria related to social justice. This overview shows a mix of technical criteria (use, cost-effectiveness), and higher-level standards linked to social choices (sustainability, justice).

Certain limitations of the study must be mentioned. The literature review was restricted to five policy fields, while other fields such as development cooperation or health were not included. To gain a better understanding on the use of criteria in the evaluation literature, it would however be interesting to include other fields in further studies such as public health – or sub-fields such as unemployment policies – as long as their particularities (e.g., a high level of standardization) are taken into account. Furthermore, for the sake of synthesis, the discussion of the results provided in this section provides only general tendencies. There is however huge diversity within each policy field, and this would merit further differentiation. In addition, the literature review focuses exclusively on the academic production around evaluations (journal articles, books, chapters), which does not give a full picture of the evaluation landscape. Had evaluation reports been included in the study, the analysis might give a partially different picture. It is possible that evaluation criteria are more transparently disclosed in reports, while not necessarily being the main focus of scientific publications. Articles in peer-reviewed journals might be less likely to report the details of evaluation projects where these do not provide information that is generalizable to broader contexts. Reports are also less accessible, because of a non-systematic publishing practice. Furthermore, the literature review was performed using the keyword "evaluation criteria." It is possible that other studies not included in the sample use different terminology to refer to criteria. In addition, in the analysis, we did not distinguish between process and efficiency evaluations because the lack of definition and discussion of evaluation criteria is an issue in both cases. The criteria used in these two types of evaluations might differ within each policy field, but this literature review aims at providing a general overview. Future studies on the use of evaluation

criteria might take this distinction into account. Finally, a possible research question that has not been addressed here is whether evaluation criteria used in the respective policy fields rather focused on micro or macro levels of analysis and whether cross-field differences are identifiable on this matter. This could be a prospect for future research.

7. Lessons learned

The foregoing analysis allows us to draw three main lessons. First, given the importance of non-intended policy effects (Birkland, 2007), evaluators and planners have great responsibility in producing knowledge beyond the linear logic of policy programs. When criteria are directly derived from program specifications, the question of the independence of evaluation might be raised (Stake, 2004). In addition to more routinized sets of criteria (e.g., effectiveness, efficiency), evaluators and planners often suggest innovative criteria as a way of shedding light on specific elements of the policy process. At the same time, these innovative criteria can turn out to be extremely numerous (often specific to each evaluation), creating what might appear to be a chaos of criteria. Thus, the existing diversity of criteria both hinders comparison between evaluations and limits the adoption of a high-level perspective on the evaluated policy, even though such a perspective is needed, especially for the evaluation of transversal policies, such as those facing the sustainability challenge.

Second, recent societal evolutions (for example, in the field of environment or land-use planning) have shown that transversal policies aimed at meeting the challenge of sustainability need to be assessed against the background of their complexity. Because these policies often imply critical choices about the future of societies, they demand an in-depth reflection on evaluation criteria. To this end, reflections on analytical dimensions (e.g., impact, relevance, acceptability, sustainability, equality) structuring the choice and use of evaluation criteria seem pertinent. In other words, the end of the technocratic illusions that characterized the second half of the twentieth century goes hand-in-hand with a necessary reexamination of evaluation procedures (how an evaluation is done) and tools (by what means it is done). As underlined by Patton (2021), while classical evaluation criteria (among others the DAC ones) can be suitable in a "business as usual situation", they remain unsatisfactorily in a situation of global social transformative action. In this context, Patton emphasizes that the inclusion of primary users in the process is crucial.

This leads to the third lesson, which is linked to the democratic dimension of evaluations. While evaluations are key to assessing the performance of public policies and programs, the dimensions against which they are judged should be subjected to debate and dialogue, rather than decided on the basis of evaluation routines or technical considerations. This raises the question of the processes through which evaluation criteria are chosen – which often constitutes a blind spot for evaluation studies. In some policy fields, a particular focus is placed on the inclusion of policy stakeholders in the criteria selection process. In those studies concerned, it is believed that assessment should be subject to a deliberative process. The participatory nature of criteria choice is of course neither necessary, nor a must, in every circumstance. However, we hold that evaluations should at least propose a transparent reflection of the way criteria were chosen, given the importance of what is at stake (e.g., nature preservation, education or welfare). This should imply an explicit and thorough reflection that starts during the early planning of each evaluation. A stronger dialogue between policy fields that have tendentially evolved in silo would enrich evaluation praxis and theory and help overcome routine bias. The superordinate categories derived from the literature review (presented in Table 1) show for instance that the use of evaluation criteria related to the democratic dimensions of policies or to social justice aspects are restricted to some policy fields (environment and social services). A reflection on the opportunity to import them in other policy evaluation fields could be done. The supra-ordinate categories also show that the use of evaluation criteria could be systematized and structured, which would go along with a more reflexive scientific attitude toward evaluation criteria. Criteria can be ordained according to whether they pertain to effects or processes (e.g., like in the environment field), or to their formal, substantive, and sustainability dimensions (e.g., land use planning field).

8. Conclusions

This study seeks to clarify the definition and use of evaluation criteria during evaluations. In fact, evaluation research has long showed that criteria must be rigorously defined and operationalized, as evaluation criteria (e.g., "quality") tend to be highly ambiguous and run therefore the risk of being subject to strategic political games (Weiss, 1982). Our study asks how the authors of evaluations understand and define evaluation criteria, what process authors follow to determine these criteria, which criteria they use, and how they use them. In the policy fields examined, the study shows that evaluation criteria are often taken for granted and are rarely subject to a reflexive definition and use. The reviewed literature refers to evaluation criteria as though the notion were based on a commonly shared understanding – a standpoint that contrasts with the observed interchanged uses of the notions of "criteria" and "indicators." Choice of evaluation criteria happens in ways that are specific to each policy field and/or study and includes inductive or deductive approaches as well as participatory or technical processes. Overall, the study reveals a great diversity of criteria, highlighting superordinate categories of criteria in each of the policy fields analyzed. Despite standardization efforts by national and international organizations, a large number of criteria exist in each policy field. Arguably, a complex and changing phenomenon such as public action is not captured at its best through a fixed and predetermined set of criteria. Although evaluations are restricted by a series of factors like time, budget, or legal requirements, it is evaluators' responsibility to be reflexive about the use of evaluation criteria. This literature on evaluation theory is unanimous about the fact that as minimal requirements, criteria should be explicitly defined, and determined before the assessment; they can be derived based on the literature, legal provisions, participatory processes, and/or existing praxis within the considered policy field. In either way evaluators might use their leeway to carefully construct criteria based on a reflection on the needs and objectives of the evaluations they conduct. Evaluations are a part society's democratic life, and therefore require transparency and accountability. Given the contemporary challenges we face, such as sustainability, there is a need to re-examine both the procedures and the means of policy evaluation. Perhaps it is high time the academic evaluation community launched a reflection on the crucial, but somewhat overlooked, subject of evaluation criteria. A cross-field dialogue on the various existing practices around evaluation criteria could make a stimulating starting point.

Funding source

none.

CRediT authorship contribution statement

François-Xavier Viallon: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Data curation, Conceptualization. **Christine Heuer:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Sigrid Haunberger:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Véronique Eicher:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Lars Balzer:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Oto Potluka:** Writing – review &

editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation. **Céline Mavrot:** Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of Competing Interest

none.

Acknowledgments

The authors want to thank their colleagues of the Swiss Evaluation Society (SEVAL) who participated in previous work of the research group.

Authors statement

We hereby confirm that all authors have participated in the conceptualization, analysis, and writing for this article. The authors have no competing interest to declare.

Appendix A. Databases included in the literature review by policy field

Policy field	Search strategy	Literature database
Social Services	"evaluation criteria" AND "social services" OR "social work" OR "welfare state"	- Sociological Abstracts (including Social Services Abstract, Sociological Abstracts) -SSOAR Social Science Open Access Repository - Field of social science -SAGE: Journals Discipline: "Social Sciences & Humanities;" Subject: "Social Work and Social Policy" - Web of Science - Google Scholar
Land-use planning	"policy evaluation" AND "evaluation criteria" AND "land-use planning"	- Google Scholar
Higher Education	"evaluation criteria" AND "higher education" AND evaluation of teaching	- PsychInfo - PSYNDEX - ERIC - SocINDEX - Education Source
Vocational Education	"evaluation criteria" AND "vocational education"	- Psychinfo - PSYNDEX - ERIC - SocINDEX - Education Source
Environment	"evaluation criteria" AND "environmental policy"	- Google.Scholar - BioMed Central; Central and Eastern European Online Library; CERN Document Server; Cochrane Library; EBSCO; ERIC; IBR online; IJBF online; IBSS; Web of Science; JSTOR; Kluwer Arbitration; MEDLINE; Oxford Reference; PLoS; ProQuest; PubMed; Scopus

Appendix B. Inclusion criteria in the literature review

Inclusion criteria
Full text in English
Published until November 2020
Full text available
Publication related to the specific policy fields
Officially published as a book, chapter or scientific journal article (no gray literature)

Appendix C. Presentation of the sample

Social services

The 34 selected sources in the "social services" policy field originate mainly from Europe (15) and the United States (13), with a few other countries also represented (6). These papers often dealt with programs or broader strategies (local service delivery networks for unaccompanied children, community violence among youth, foster care, case management, housing for the elderly, after-school care programs), and have sometimes developed indicators (e.g., positive aging indicators or child-protection assessment). In terms of their methods, the papers consist of theoretical discussions (referring to programs and evaluation models) (9), systematic literature reviews (6), evaluations and empirical studies using a qualitative approach, sometimes based on mixed-methods designs (combining different methods such as interviews, observations, document review, focus groups and case studies) (15), and studies using quantitative methods (surveys, secondary data analyses, factorial survey, etc.) (4).

Land-use planning

In comparison to the other policy fields, the field of land-use planning was searched using the additional term "policy evaluation". This term was added because of a first search without the additional term that proved unsuccessful. In fact, the first search produced highly heterogeneous results without clear connection to evaluation in the field. Among the results, most papers stem from European research (31), followed by North American

(11), Asian (9), and South American research (1). Of the 52 articles considered for analysis, 29 report on an empirical policy evaluation, eight focus on theoretical considerations around policy evaluation, six deal with methodological issues of evaluation in the field, and nine consider empirical questions together with theoretical or methodological aspects of policy evaluation. Of the 27 quantitative papers, 19 use multi-criteria decision analysis as a method for the evaluation of land-use plan implementation. Of the 18 qualitative papers, nine use case study methodology to evaluate land-use plan implementation.

Teaching in higher education

Of the sample of 22 sources in the field of teaching in higher education, half are from the USA, while approximately a fifth are from Europe. Half are written by authors in the field of educational sciences and a quarter by authors from neighboring disciplines (such as psychology and sociology). About a third of the papers are theoretical, and of the empirical papers (those based on actual evaluation experience), 80 % use quantitative methods and almost two-thirds focus on the implementation of an evaluation, while one-third is concerned more with research on how to do evaluations (e.g., should indicators be weighted or not?).

Vocational education training

While 101 articles were identified via a general search, only 24 articles matched the inclusion criteria. Most of the papers excluded focused on assessment in the sense of measurement of competencies, rather than on evaluation. The former is important (especially in education) and covers a large body of research, but this is not relevant to our study. In the remaining sample of 24 articles, about two-thirds of the publications are from Europe, about one quarter from the USA and just a few are from Australia and Africa – which is not surprising given the importance of vocational education and training (VET) in many European countries. Almost two-thirds are written as research texts, while the others have a practical perspective. Just under 17 % of the papers are theoretical, and of the empirical papers, more than three- quarters use quantitative methods. The journals are fairly diverse, with no journal featuring more than two articles. Nearly 80 % of journals are located in the educational field, with only a few in specific domains such as counseling or rehabilitation. Around 17 % are specific evaluation journals; this is important because most of the texts use field-specific logic and language, which is sometimes quite different from the evaluation language.

Environment

Of the 47 articles selected, 22 are empirical case studies (often using multi-criteria analysis), and 20 are theoretical contributions, of which five are methodological (sometimes with illustrative applications), and three are non-systematic literature reviews. Empirical case studies thus dominate the sample, followed by theory-building pieces. These case studies ranged from evaluations of single policies or programs to multiple case studies of up to 13 evaluation units. Articles focus on various themes including wetlands protection, wastewater management, forest protection, biodiversity, air pollution, and energy policy. In terms of geography, the case studies cover the five continents, with a predominance of European cases.

References

Abma, T. A. (2000). Stakeholder conflict: A case study. *Evaluation and Program Planning, 23*(2), 199–210. https://doi.org/10.1016/S0149-7189(00)00006-9

Alkin, M.C. (2013). Evaluation Roots - A Wider Perspective of Theorists' Views and Influences (2nd ed.). Los Angeles: Sage Publications.

Alkin, M. C., Vo, A. T., & Christie, C. A. (2012). The Evaluator's Role in Valuing: Who and With Whom. *New Directions for Evaluation, 133*, 29–41. https://doi.org/10.1002/ev

Amin, M. E. (2002). Six Factors of Course and Teaching Evaluation in a Bilingual University in Central Africa. *Assessment Evaluation in Higher Education, 27*(3), 281–291. https://doi.org/10.1080/0260293022013863

Arthur, W., Tubré, T., Paul, D. S., & Edens, P. S. (2003). Teaching Effectiveness: The relationship between reaction and learning evaluation criteria. *Educational Psychology, 23*(3), 275–285. https://doi.org/10.1080/0144341032000060110

Baartman, L., Gulikers, J., & Dijkstra, A. (2013). Factors Influencing Assessment Quality in Higher Vocational Education. *Assessment Evaluation in Higher Education, 38*(8), 978–997.

Baker, D. F., Neely, W. P., Prenshaw, P. J., & Taylor, P. A. (2015). Developing a Multi-Dimensional Evaluation Framework for Faculty Teaching and Service Performance. *Journal of Academic Administration in Higher Education, 11*(2), 29–41.

Ball, M. (1988). Evaluation in the voluntary sector. London: Forbes Trust.

Balzer, L., Frey, A., & Nenniger, P. (1999). Was ist und wie funktioniert Evaluation? *Empirische Pädagogik, 13*(4), 393–414.

Barron, W. F., & Ng, G. T. L. (1996). An assessment methodology for environmental policy instruments: An illustrative application to solid wastes in Hong Kong. *Journal of Environmental Management, 48*(3), 283–298. https://doi.org/10.1006/jema.1996.0078

Beckers, J., Dolmans, D. H. J. M., & van Merriënboer, J. J. G. (2019). PERFLECT: Design and Evaluation of an Electronic Development Portfolio Aimed at Supporting Self-Directed Learning. *TechTrends: Linking Research and Practice to Improve Learning, 63*(4), 420–427.

Bellamy, J. A., Walker, D. H., McDonald, G. T., & Syme, G. J. (2001). A systems approach to the evaluation of natural resource management initiatives. *Journal of Environmental Management, 63*(4), 407–423. https://doi.org/10.1006/jema.2001.0493

Biggs, J. B. (1993). From Theory to Practice: A Cognitive Systems Approach. *Higher Education Research Development, 12*(1), 73–85.

Birkland, T. A. (2007). Agenda-Setting in Public Policy. In F. Fischer, & G. J. Miller (Eds.), *Handbook of Public Policy Analysis: Theory, Politics and Methods*. CRC Press.

Campbell, D. T. (1991). Methods for the Experimenting Society. *Evaluation Practice, 12*(3), 223–260. https://doi.org/10.1177/109821409101200304

Chipman, R., Wells, S. J., & Johnson, M. A. (2002). The meaning of quality in kinship foster care: Caregiver, child, and worker perspectives. *Families in Society, 83*, 508–520.

Clarke, N. (2001). The impact of in-service training within social services. *British Journal of Social Work, 31*, 757–774.

Custer, R. L., Ruhland, S. K., & Stewart, B. R. (1997). Assessing Tech Prep Implementation. *Journal of Vocational and Technical Education, 13*(2), 23–35.

Daniel, M., Cargo, M., Marks, E., Paquet, C., Simmons, D., Williams, M., … O'Dea, K. (2009). Rating health and social indicators for use with indigenous communities: A tool for balancing cultural and scientific utility. *Social Indicators Research, 94*, 241–256.

Davidson, J.E. (2005a). Evaluation Methodology Basics. The Nuts and Bolts of Sound Evaluation. Thousand Oaks: Sage.

Davidson, J. E. (2005b). Evaluation Values and Criteria Checklist 2001. In S. Mathison (Ed.), *Encyclopedia of Evaluation*. Thousand Oaks: Sage.

De Maeyer, S., van den Bergh, H., Rymenans, R., Van Petegem, P., & Rijlaarsdam, G. (2010). Effectiveness Criteria in School Effectiveness Studies: Further Research on the Choice for a Multivariate Model. *Educational Research Review, 5*(1), 81–96.

De Vos, M. E., Baartman, L. K. J., Van Der Vleuten, C. P. M., & De Bruijn, E. (2019). Exploring How Educators at the Workplace Inform Their Judgement of Students' Professional Performance. *Journal of Education and Work, 32*(8), 693–706.

Del Carmen Bas, M., Tarantola, S., Carot, J. M., & Conchado, A. (2017). Sensitivity Analysis: A Necessary Ingredient for Measuring the Quality of a Teaching Activity Index. *Social Indicators Research, 131*(3), 931–946. https://doi.org/10.1007/s11205-016-1297-2

Dickinson, P., & Adams, V. (2017). Values in evaluation - The use of rubrics. *Evaluation and Program Planning, 65*, 113–116. https://doi.org/10.1016/j.evalprogplan.2017.07.005

Dunkin, M. J. (1995). Concepts of Teaching and Teaching Excellence in Higher Education. *Higher Education Research Development, 14*(1), 21–33. https://doi.org/10.1080/0729436950140103

EC. (2019). The European Green Deal, available at: ⟨https://eur-lex.europa.eu/resource.html?uri=cellar:b828d165-1c22-11ea-8c1f-01aa75ed71a1.0002.02/DOC_1&format=PDF).

Eeckhout, M., Jong, N. de, & Valk, P. de (1996). Programme aid evaluation: Dutch practice and a case for an economic and institutional approach at multi-donor level. *IDS Bulletin, 27*, 79–87.

Essack, S., Naidoo, I., Oosthuizen, F., Bodenstein, J., Brysiewicz, P., & Suleman, F. (2012). Quality Teaching and Learning in the Health Sciences. *South African Journal of Higher Education, 26*(5), 960–972. https://doi.org/10.20853/26-5-203
Garey, A. I. (2002). Social domains and concepts of care: Protection, instruction, and containment in after-school programs. *Journal of Family Issues, 23*, 768–788.
Gibney, S., Sexton, E., & Shannon, S. (2019). Measuring what matters: achieving consensus on a positive aging Indicator set for Ireland. *Journal of Aging Social Policy, 31*, 234–249.
Gilgun, J. F. (1988). Decision-making in interdisciplinary treatment teams. *Child Abuse Neglect, 12*, 231–239.
Gleason, S. E. (1986). Inmate Attitudes Toward Vocational Training: A Case Study of Vocational Training Students in the State Prison of Southern Michigan. *Journal of Offender Counseling, Services Rehabilitation, 10*(4), 49–60. https://doi.org/10.1300/J264v10n04_05
Goulder, L. H., & Parry, I. W. H. (2008). Instrument choice in environmental policy. *Review of Environmental Economics and Policy, 2*(2), 152–174. https://doi.org/10.1093/reep/ren005
Gunningham, N., & Young, M. D. (1997). Toward Optimal Environmental Policy: The Case of Biodiversity. *Ecology Law Quarterly, 14*(243), 55. Retrieved from ⟨http://scholarship.law.berkeley.edu/cgi/viewcontent.cgi?article=1538&context=elq⟩.
Gysen, J., Bruyninckx, H., & Bachus, K. (2006). The Modus Narrandi: a methodology for evaluating effects of environmental policy. *Evaluation, 12*(1), 95–118. https://doi.org/10.1177/1356389006064176
Hanberger, A., Wimelius, M. E., Ghazinour, M., Isaksson, J., & Eriksson, M. (2016). Local service-delivery networks for unaccompanied children in Sweden: Evaluating their effectiveness. *Journal of Social Service Research, 42*, 675–688.
Harth, H., & Hemker, B. T. (2013). On the Reliability of Vocational Workplace-Based Certifications. *Research Papers in Education, 28*(1), 75–90.
Heuer, C. (2017). Verwendung von Bewertungskriterien in den externen Evaluationen der Bundesverwaltung: Ergebnisse einer Forschungsstudie. *LeGes, 2*, 327–345.
Hideg, M. (2019). Understanding the evaluation of public policy. evaluation of the romanian national anti-drug strategy. *Social Research Reports, 11*, 81–89.
Höhns, G. M. (2017). Pedagogic Practice in Company Learning: The Relevance of Discourse. *Journal of Vocational Education and Training, 70*(2), 313–333.
Holmes, J., & Clark, R. (2008). Enhancing the use of science in environmental policy-making and regulation. *Environmental Science and Policy, 11*(8), 702–711. https://doi.org/10.1016/j.envsci.2008.08.004
House, E.R., & Howe, K.R. (1999). Values in Evaluation and Social Research. Thousand Oaks: Sage Publications.
Kagle, J. D. (1979). Evaluating social work practice. *Social Work, 24*, 292–296.
Kastein, M. R., Jacobs, M., van der Hell, R. H., Luttik, K., & Touw-Otten, F. W. M. M. (1993). Delphi, the issue of reliability: a qualitative Delphi study in primary health care in the Netherlands. *Technological Forecasting and Social Change, 44*, 315–323.
Khaleel, I. A. (1988). The Spiral-Interactive Program Evaluation Model. *Educational Technology, 28*(5), 43–46.
Kim, S. H. (2007). Evaluation of negative environmental impacts of electricity generation: Neoclassical and institutional approaches. *Energy Policy, 35*(1), 413–423. https://doi.org/10.1016/j.enpol.2005.12.002
Kirkpatrick, D. (1976). Evaluation of training. In In. R. L. Craig (Ed.), *Training and Development Handbook: A Guide to Human Resource Development* (pp. 301–319). New York: McGraw-Hill.
Kirkpatrick, D. (1996). Great ideas revisited. *Training Development, 50*(1), 54–59. Retrieved from ⟨http://ezaccess.libraries.psu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a2h&AN=9602066395&site=ehost-live%5Cnhttp://content.ebscohost.com/ContentServer.asp?T=P&P=AN&K=9602066395&S=R&D=a2h&EbscoContent=dGJyMMTo50Sep7U4yNfsOLCmr0ye⟩.
Knoepfel, P., Larrue, C., Varone, F., & Hill, M. (2007). Public policy analysis. Public Policy Analysis. Bristol: Policy Press.
Kollöffel, B., & de Jong, T. (2016). Can Performance Feedback during Instruction Boost Knowledge Acquisition? Contrasting Criterion-Based and Social Comparison Feedback. *Interactive Learning Environments, 24*(7), 1428–1438.
Kunseler, E. M., & Vasileiadou, E. (2016). Practising environmental policy evaluation under co-existing evaluation imaginaries. *Evaluation, 22*(4), 451–469. https://doi.org/10.1177/1356389016668099
Lacouture, A., Breton, E., Guichard, A., & Ridde, V. (2015). The concept of mechanism from a realist approach: a scoping review to facilitate its operationalization in public health program evaluation. *Implementation Science, 10*, 1–10.
Langemeyer, J., Gómez-Baggethun, E., Haase, D., Scheuer, S., & Elmqvist, T. (2016). Bridging the gap between ecosystem service assessments and land-use planning through Multi-Criteria Decision Analysis (MCDA). *Environmental Science and Policy, 62*, 45–56. https://doi.org/10.1016/j.envsci.2016.02.013
Langer, A., Eurich, J., & Güntner, S. (2019). Innovation, Quality and Evaluation. *In Innovation in Social Services* (pp. 41–48). Wiesbaden: Springer.
Linder, S. H., & Peters, B. G. (1998). The study of policy instruments: four schools of thought. In In. B. G. Peters, & F. K. M. van Nispen (Eds.), *Public policy instruments: evaluating the tools of public administration* (pp. 33–45). Cheltenham: Edward Elgar.
Machi, L.A., & McEvoy, B.T. (2016). The literature review: Six steps to success. London: Corwin Press.
Mazmanian, D.A., & Sabatier, P.A. (1983a). Implementation and public policy. Glenview Ill: Scott Foresman.
Mazmanian, D.A., & Sabatier, P.A. (1983b). Implementation and public policy Glenview. IL Scott Foresman.
Min, L., & Huilan, X. (2020). Comparative analysis of long-term care quality for older adults in China and Western countries. *Journal of International Medical Research, 48*, 0300060519865631.
Moro, G., Cassibba, R., & Costantini, A. (2007). Focus groups as an instrument to define evaluation criteria: The case of foster care. *Evaluation, 13*, 340–357.
Munda, G., Nijkamp, P., & Rietveld, P. (1994). Qualitative multicriteria evaluation for environmental management. *Ecological Economics, 10*(2), 97–112. https://doi.org/10.1016/0921-8009(94)90002-7
Muñoz Gielen, D., & Mualam, N. (2019). A framework for analyzing the effectiveness and efficiency of land readjustment regulations: Comparison of Germany, Spain and Israel. *Land Use Policy, 87*, Article 104077. https://doi.org/10.1016/j.landusepol.2019.104077
OECD/DAC. (2021). Applying Evaluation Criteria Thoughtfully. Paris: Organisation for Economic Co-operation and Development. https://doi.org/10.1787/543e84ed-en.
OECD. (1991). Principles for evaluation of development assistance. Paris: Organisation for Economic Co-operation and Development. ⟨https://www.oecd.org/dac/evaluation/2755284.pdf⟩ [last accessed 5 August 2022].
OECD (2021), Government at a Glance 2021, OECD Publishing, Paris, https://doi.org/10.1787/1c258f55-en.
Oliveira, V., & Pinho, P. (2010). Evaluation in urban planning: Advances and prospects. *Journal of Planning Literature, 24*(4), 343–361. https://doi.org/10.1177/0885412210364589
Patton, M. Q. (2021). Evaluation Criteria for Evaluating Transformation: Implications for the Coronavirus Pandemic and the Global Climate Emergency. *American Journal of Evaluation, 42*(1), 53–89. https://doi.org/10.1177/1098214020933689
Portney, P. R., & Stavins, R. N. (1994). Regulatory review of environmental policy: The potential role of health-health analysis. *Journal of Risk and Uncertainty, 8*(1), 111–122. https://doi.org/10.1007/BF01064089
Pratama, A., & Setiawan, I. (2018). Program evaluation coaching on abandoned children who drop out of school in PPSBR Makkareso in Maros. *Journal of Physics: Conference Series, 1028*, Article 012149.
Ressing, M., Blettner, M., & Klug, S. J. (2009). Systematische Übersichtsarbeiten und Metaanalysen. *Dtsch Arztebl Int, 106*, 456–463.
Richards, K. R. (2000). Framing environmental policy instrument choice. *Duke environmental Law and Policy Forum, 10*(2), 221–286.
Rieper, O., & Mayne, J. (1998). Evaluation and public service quality. *Scandinavian Journal of Social Welfare, 7*, 118–125.
Romainville, M. (1999). Quality Evaluation of Teaching in Higher Education. *Higher Education in Europe, 14*(3), 415–424.
Rossi, P.H., Lipsey, M.W., & Henry, G.T. (2019). Evaluation: A Systematic Approach (8.). Thousand Oaks: Sage.
Rycroft, R. W. (1978). Selecting Policy Evaluation Criteria: Toward a Rediscovery of Public Administration. *The American Review of Public Administration, 12*(2), 87–98.
Sager, F., & Mavrot, C. (2021). Participatory vs Expert Evaluation Styles. In M. Howlett, & J. Tosun (Eds.), *Routledge Handbook of Policy Styles* (pp. 395–407). London: Routledge.
Salamon, L. (2002). The tools of government: a guide to the new governance. In In. L. Salamon (Ed.), *The Tools of Government: A Guide to the New Governance* (pp. 1–48). Oxford: England: Oxford University Press.
Schmidt-Posner, J., & Jerrell, J. M. (1998). Qualitative analysis of three case management programs. *Community Mental Health Journal, 34*, 381–392.
Scriven, M. (1980). The Logic of Evaluation. California: Edg. Press.
Scriven, M. (1991). Evaluation Thesaurus (4th ed.). Newbury Park: Sage Publications.
Scriven, M. (2007). The Logic of Evaluation, 1–16. URL: ⟨https://scholar.uwindsor.ca/cgi/viewcontent.cgi?article=1390&context=ossaarchive⟩ [last accessed 5 August 2022].
Scriven, M. (2015). Key Evaluation Checklist. URL: ⟨http://www.michaelscriven.info/images/MS_KEC_8-15-15.doc⟩ [last accessed 5 August 2022] S.
Shadish, W.R., Cook, T., & Leviton, L.C. (1991). Foundations of program evaluation: Theories of practice. Thousand Oaks: Sage Publications.
Song, I. H., Soskolne, V., Zuojian, Z., Browne, T., & Wong, J. (2019). Global Health Social Work. In In. S. Gehlert, & T. Browne (Eds.), *Handbook of Health Social Work* (pp. 71–91). Hoboken, NJ: Wiley.
Sørensen, R., & Bay, A. (2002). Competitive tendering in the welfare state: perceptions and preferences among local politicians. *Scandinavian Political Studies, 25*, 357–384.
Sritanyarat, D. (2014). Development of Theoretical Based Multidimensional Learners' Evaluation of Higher Education in Thailand: A Case Study of the University in Graduate Level. *NIDA Development Journal, 54*(4), 17–56.
Stake, R.E. (1995). The art of case study research. Thousand Oaks: Sage.
Stake, R.E. (2004). Standards-based and responsive evaluation. Thousand Oaks: Sage.
Stockmann, R., & Meyer, W. (2010). Evaluation - eine Einführung. Opladen & Farmington Hills: Verlag Barbara Budrich.
Stufflebeam, D.L. (2001). Evaluation Values and Criteria Checklist. URL: ⟨https://wmich.edu/sites/default/files/attachments/u350/2018/values-criteria-stufflebeam.pdf⟩ [last accessed 5 August 2022].
Stufflebeam, D.L., & Zhang, G. (2017). The CIPP Evaluation Model: How to Evaluate for Improvement and Accountability. New York: Guilford Press.
Tavory, I., Timmermans, S. (2014). Abductive Analysis: Theorizing Qualitative Research. Chicago, London: Chicago University Press.
Thomas, P., & Palfrey, C. (1996). Evaluation: stakeholder-focused criteria. *Social Policy Administration, 30*, 125–142.
Van Den Hove, S. (2000). Participatory approaches to environmental policy-making: The European Commission Climate Policy Process as a case study. *Ecological Economics, 33*(3), 457–472. https://doi.org/10.1016/S0921-8009(99)00165-2
Vedung, E. (1998). Policy instruments: typologies and theories. In In. M.-L. Bemelsman-Videc, R. Rist, & E. Vedung (Eds.), *Carrots, sticks and sermons. Policy instruments and their evaluation* (pp. 21–58). New Brunswick and London: Transaction publishers.

Voisin, D. R., & Berringer, K. R. (2015). Interventions targeting exposure to community violence sequelae among youth: A commentary. *Clinical Social Work Journal, 43*, 98–108.
Waterhouse, L., & Carnie, J. (1992). Assessing child protection risk. *The British Journal of Social Work, 22*, 47–60.
Weiss, C. H. (1982). Policy Research in the Context of Diffuse Decision Making. *The Journal of Higher Education, 53*(6), 619–639.
Weiss, C. H. (1993). Where Politics and Evaluation Research Meet. *Evaluation Practice, 14* (1), 93–106.
Weiss, C.H. (1998). Evaluation - methods for studying programs and policies (2nd ed.). Upper Saddle River: Prentice Hall.
Widmer, T., & De Rocchi, T. (2012). Evaluation: Grundlagen, Ansätze und Anwendungen. Zürich: Rüegger.
Zamfir, E. (2017). Quality of life and social justice in Romania: Measuring quality ocrf life. *Revista Dèleľött Cercetare Şi Intervenţie Socială, 58*, 34–53.

Céline Mavrot is Assistant Professor at the University of Lausanne, Switzerland. She specializes in policy evaluation and comparative policy analysis, with a focus on health policy and public controversies. She regularly conducts policy evaluations for public agencies and non-governmental organizations and teaches evaluation in several curricula. Her work has been published in numerous journals. In 2020, she won the Award of the Swiss Evaluation Society together with Susanne Hadorn and Fritz Sager.

Oto Potluka is a senior researcher at the Center for Philanthropy at the University of Basel, Switzerland. His main concerns are evaluations of impact of public expenditure programs. He specializes on the EU Cohesion policy and regional development, including the role of the civil society in development policies. He is a member of the European Evaluation Society, the Swiss Evaluation Society (SEVAL), and the Czech Evaluation Society.

Lars Balzer is head of the Evaluation Unit at the Swiss Federal University for Vocational Education and Training SFUVET in Zollikofen (Switzerland). In 1998, he graduated in psychology. He worked as a research assistant at the Centre for Educational Research at the University of Koblenz-Landau (Germany), with a doctorate in psychology in 2005. Since 2005 he works at the SFUVET. In 2017, he become a Professor at the SFUVET. His primary interests include evaluations of reform and innovation projects in vocational education and training, and research on evaluations.

Véronique Eicher is project manager at the Evaluation Unit of the Swiss Federal University for Vocational Education and Training. She is responsible for the evaluation of teaching, as well as conducting internal and external evaluation projects. Additionally, she teaches courses in evaluation.

Sigrid Haunberger is a professor at the Institute of Social Management, ZHAW Zurich University of Applied Sciences. Her research interests are in the areas of social structure analysis, career and major choice motives, impact evaluations in health and social care, volunteer management and volunteer engagement, research methods and survey research, social work in the correctional system.

Christine Heuer was project manager at the Evaluation and Research Service of the Swiss Federal Office of Public Health. She performed the process management for evaluation projects and assured the quality of the evaluation results.

François-Xavier Viallon is a project manager at the Swiss Federal University for Vocational Education and Training, Zollikofen. His research interests include public policy implementation, natural resource management and property governance.