Article

Evaluation
16(3) 249–262

DOI: 10.1177/1356389010370262
http://evi.sagepub.com
SAGE

Criteria and Theory in the Evaluation of Organizations

Barbara Befani
Independent Evaluation Consultant, Italy

Abstract
This article argues that evaluation-specific logic, which involves the identification of relevant criteria and performance standards for the evaluand (considered as a whole or in its single dimensions), can benefit from research on the causal chains leading from programme inputs to the final outcomes; and on the contexts and mechanisms responsible, which are normally investigated when adopting a theory-based or realistic evaluation approach. The arguments presented are both theoretical – how explanations can be used to redefine the evaluand in order to improve criteria selection – and applied. The latter are taken from the evaluation of two organizations: Italian universities and the Food and Agricultural Organization of the United Nations (FAO).

Keywords
criteria, evaluation-specific logic, international organizations, theory, university

The combination of evaluation-specific logic and theory-based evaluation is rare. The former involves establishing the criteria by which the merit of an evaluand will be judged, usually through the following steps:

- the selection of a set of evaluation criteria pertaining to the evaluand;
- the identification of a set of evaluation standards, e.g. what constitutes good, fair, poor, excellent, etc. for each criterion or a combination of them;
- rating the performance of the evaluand on the criteria and
- synthesizing the findings into an overall grade, rank, or score.

On the other hand, theory-based evaluation involves discovering the inner workings of a programme in terms of social mechanisms being activated in and among the stakeholders that are responsible for obtaining a certain effect or result. Apparently, the two approaches have different aims and are characterized by irreconcilable differences. According to E. Jane Davidson (2005), it is a 'common view' that 'the use of evaluation logic and methodology is somehow the antithesis of theory-based evaluation'. While Davidson is surprised ('in fact, this is one of the most powerful blends possible'), the father of the evaluation-specific logic states that theories are 'a luxury for the

Corresponding author:
Barbara Befani, Independent Evaluation Consultant, Italy.
Email: befani@gmail.com

evaluator, since they are not even essential for explanations, and explanations are not essential for 99% of all evaluations' (Scriven, 1991).

This article aims to shed light on the controversy by showing that the conflict is only apparent. In the absence of an explicit attribution of merit or value, evaluative research is not immediately distinct from ordinary social research; however, merit or value attribution often require knowledge about the mechanisms responsible for the programme outcomes. The latter is argued by stressing that evaluation criteria for an evaluand are chosen not only from its 'static' properties, but also from its 'dynamic' features; they are also identified through observation of the evaluand's interaction with the outside world and the changes that the external environment causes to the evaluand's inner functioning. An evaluand's reaction to an environment or context cannot usually be overlooked when deciding 'where to look' in the process of establishing which of its dimensions will be relevant for the analysis. Failing to conceive of an evaluand 'dynamically' can lead to the selection of the wrong set of criteria and thus impair the validity of the exercise from the start, long before choosing which method of data analysis to employ.

This thesis finds support in the evaluation of organizations. When deciding which dimensions/criteria are going to be relevant for understanding the merits of an organization or some of its units, failing to explicitly construct a 'causal chain' connecting at least some of the functions being performed and analysing their interrelations and interactions can be very risky. Perhaps more than other evaluands, organizations are living organisms that react to changes in the outside world according to their own internal laws; it is thus necessary to register these reactions in order to control for them, or take them into account during the evaluation. It does not make sense to benchmark a final outcome or product or service, comparing it to what was produced in the past or to what other organizations are producing, without knowing how the operational context has changed or the conditions in which the different organizations are operating. The different contexts give rise to specific internal processes and workflows, the differences among which cannot be ignored when focusing on what changes the organization needs to enact to improve performance.

Theoretical Elements of the Apparent Conflict

According to the founding father of evaluation-specific logic, theories do not help evaluation because they 'are not even essential for explanation, and explanations are not essential for 99% of evaluations' (Scriven, 1991). The two relationships to be explored are thus the one between theory and explanation, and the one between explanation and evaluation. The argument that evaluation does need explanations builds on the idea that when the evaluand is a social science concept it needs to be defined using a causal, ontological and realist approach to concept definition;[1] which involves 'ascertaining the constitutive characteristics of a phenomenon that have central causal powers' (Goertz, 2006).

Law vs Explanation

In its traditional definition, a scientific theory is a body of laws, whereby a law is a 'general account of a field of phenomena, generating at least explanations and sometimes also predictions and generalizations' (Scriven, 1991). A scientific theory in its 'standard' form is thus expressed as a law asserting that, 'given certain initial conditions, an event of a given type (the cause) will always produce an event of some other type (the effect)' (Elster, 1998). In other words, laws are statements of the kind 'if A, then always B'. Because they have high explanatory power, laws can be useful for the evaluator; however, as the latter's field is usually the social sciences, laws are also

in short supply for her/him to use as there are no law-like generalizations in the social sciences. 'Explanation by laws is better – but also more difficult, usually too difficult' (Elster, 1998). Social scientists have long rejected attempts to develop general systems of sociological theory and advocated instead that sociological theory should deal with social mechanisms (Hedström and Swedberg, 1998). Theory is indeed not essential for explanation: while theories explain, explanations do not necessarily enjoy the level of generalization that theories require. Mechanisms are just 'sometimes true theories' (Coleman, 1964) and take the form 'if A, then sometimes B' (Elster, 1998).

Social scientists who are also evaluators use the word theory in a way that is closer to the concept of explanation than to the concept of law (Donaldson, 2007; Pawson, 2006; Pawson and Tilley, 1997;Weiss, 1972, 1997). Some state that their definition of theory is 'broader than the conventional definition of scientific theory', and define it as 'a set of interrelated assumptions, principles, and/or propositions to explain . . . social actions', that is 'frequently . . . unsystematic' and 'explains how the program is supposed to work' (Chen, 1990). Others provide even less stringent definitions of theory: 'a set of propositions regarding what goes on in the black box during the transformation of input to output; that is, how a bad situation is transformed into a better one through treatment inputs' (Lipsey, 1993).

Explanation vs Evaluation

The idea that explanations are not essential for the vast majority of evaluations has been criticized by many authors. For example, the statement 'one does not need to know anything at all about electronics to evaluate electronic typewriters' (Scriven, 1991) has been criticized by highlighting that the assumption on which it rests is that 'the theory which would be used would be drawn from electronics', while in fact it 'relates to the mechanisms through which the introduction of electronic typewriters may generate improvements in outcome in the contexts in which they are placed' (Pawson and Tilley, 1997).

In their breakthrough book of 1997, Pawson and Tilley state that the main objective of evaluation research is the discovery or refinement of the social science theory underlying the programme being evaluated. In their approach, programme theory is expressed in the renowned Context–Mechanism–Outcome (CMO) form, in which the policy outcome is seen as the result of an underlying causal force (the mechanism) operating in a given context. In other words, a specific policy outcome is explained by the actions, reasoning, or choices made by stakeholders embedded in a given resource structure, defined by specific opportunities and constraints of varying nature (social, legal, economic, relational, geographic, cultural and so on) (Pawson and Tilley, 1997).

The idea that explanations are necessary in evaluative research is also central to Carol Weiss's work; the author sees the reconstruction of the causal chain linking the programme's inputs to the programme's outcomes as the core of the evaluative effort. 'What ideas and assumptions link the program's inputs to the attainment of the desired ends?'. While Weiss uses the word theory, she distances herself from the law-like notion of theory, specifying that by theory, she does not mean 'anything highbrow or multi-syllabic', but rather 'the set of beliefs that underlie action'; and that the theory at hand 'doesn't have to be uniformly accepted [nor] right', but only needs to be 'a set of hypotheses upon which people build their program plans' and 'an explanation of the causal links that tie program inputs to expected program outputs' (Weiss, 1972).

While other evaluators provide similar definitions of programme theory where explanations and causal chains are crucial targets – 'a plausible and sensible model of how a program is supposed to work' (Bickman, 1987); 'a chain of causal assumptions linking program resources, activities, intermediate outcomes, and ultimate goals' (Wholey, 1987) – Weiss offers further insight on the concept

of mechanism which is reminiscent of realistic evaluators: 'the operative mechanism of change isn't the program activities per se but the response that the activities generate' (Weiss, 1972).

Defining the Evaluand (or a Social Science Concept)

It would seem that evaluators are very interested in explanations. This paragraph outlines and clarifies the modalities in which explanations can serve the purpose of merit and value establishment from a theoretical point of view. In order to select the relevant criteria to evaluate an evaluand, one first needs a definition of the evaluand: in other words, one needs to be sure that a given, selected criterion is a feature or property of the evaluand in the first place. An evaluand is a concept and as such needs to be carefully defined; failure to define it correctly might lead to incorrect assumptions about the object of the evaluation, hindering the validity and robustness of the exercise from the start.

How do we define a concept? In probative logic, 'the relation between concepts and criteria for them replaces the relation in classical logic between concepts and their defining features' (Scriven, 1991). In other words, criteria are what are used in probative logic to define a concept and can be roughly identified with its defining features (Scriven, 1959). If the concept is an evaluand, then the evaluation criteria must also be its defining features. This idea is normally presented to argue that a set of criteria is sufficient to define an evaluand by scientific standards (Scriven, 1995); however, it also shows that, whatever criteria are picked, they must be part of the object definition.

When one evaluates an apple in terms of its flavour, it is because: (a) the flavour is relevant to her/him but also because (b) the apple *has* a flavour; the flavour is a property of the apple. The fact that the flavour 'belongs' to the apple is a required precondition in order for it to be picked as a criterion in the evaluation of an apple. When one evaluates a programme in terms of its impact on local development, it is because (a) the programme's impact on local development is relevant to the evaluator but also because (b) the programme is defined as something that can have an impact on local development, otherwise it would not make sense to include impact in the evaluation criteria. One does not include flavour among the criteria for the programme, not because one does not value a good apple flavour, but simply because flavour is not one of the defining features of the programme. Similarly, one does not define an apple as having a potential impact on local development, but rather on one's need for a balanced diet (criterion: inclusion of specific nutrients), one's hunger at a given moment (criterion: ability to make one feel less hungry) or the satisfaction of one's gluttony (criterion: flavour). Both inclusion of specific nutrients, ability to make one feel less hungry and flavour are properties of the apple, not of the programme for local development. Therefore, one can only select them for the apple evaluation, because they are defining features of that evaluand; while for other evaluands one needs to use different criteria because they have different defining features.

In doing an evaluation, the evaluator needs to select the criteria about the evaluand that are most value-relevant (to her, to the client, to the beneficiaries, etc.); however, not all the defining features of the evaluand are going to be value-relevant: for example, one likely will not be interested in the colour of the apple's stalk or in the colour of chairs that are used in the implementation of a programme. Nonetheless, all evaluative criteria need to be part of the defining features of the evaluand, even those that are left out or considered unimportant (the stalk must belong to the apple and the programme at some point needed a set of coloured chairs). In brief, defining the evaluand is crucial for proper criteria selection. In mathematical terms, the set of relevant criteria for an evaluand are a subset of its defining features or properties (see Figure 1).

The problem with failing to find explanations why the evaluand assumes a certain behaviour or presents specific results (the core objective of theory-based evaluation) lies in the fact that – without

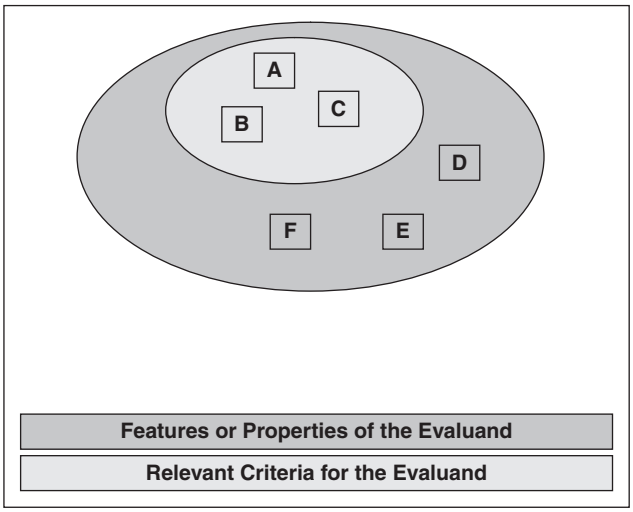

Figure 1. The relevant criteria as a subset of the evaluand's defining features

the information deriving from those explanations – our definition of the evaluand might not be complete or correct. There might be some criterion, some dimension of the evaluand that we have ignored which, had we known about, would have been considered a relevant feature. Discovering why an evaluand behaves the way it does has the potential to reshuffle the distribution of our 'value stock' among the criteria, adding new ones or changing the importance assigned to each of them. For example, if we discover that a certain result has been brought about in a way that infringes some regulations that should not have been transgressed, that would most likely reshuffle our set of criteria to explicitly include respect of those regulations.

Causal factors or mechanisms are not only causes of evaluation criteria; sometimes they are also evaluation criteria themselves, and for the same evaluand. That is because sometimes an evaluand cannot easily be distinguished from its causes. In its definition, the separation between the object itself and the mechanisms through which it interacts with an external environment sometimes is not clear cut. When choosing a criterion for an evaluand, sometimes we are not interested in that criterion per se, but in what it does; in its dynamic features, rather than its static properties. Back to the apple example, we appreciate its flavour because of the effect it has on our senses, and we value its nutritional properties because of what its specific nutrients do to our body. We are not interested in the apple as a static object but as a dynamic one that interacts with an environment we are part of; it is not so much the evaluand in itself, but rather what it can do, to us and our surroundings.

In his book on social science concepts, Gary Goertz emphasizes the importance of mechanisms in the definition of concepts:

> . . . the ontological theory expounded by the concept focuses on the concept's internal structure and its constituent parts, but that analysis is intimately related to how the object as a whole interacts, usually in a causal way, with its environment. We tend to identify as core dimensions those that have causal powers when the object interacts with the outside world . . . One cannot neatly separate the ontology of a concept from the role it plays in causal theories and explanations . . . Much of what good ontology entails is an analysis of those properties which have causal powers and which are used in causal explanations and mechanisms. The atomic structure of copper explains how it acts in many situations, e.g., its conductivity,

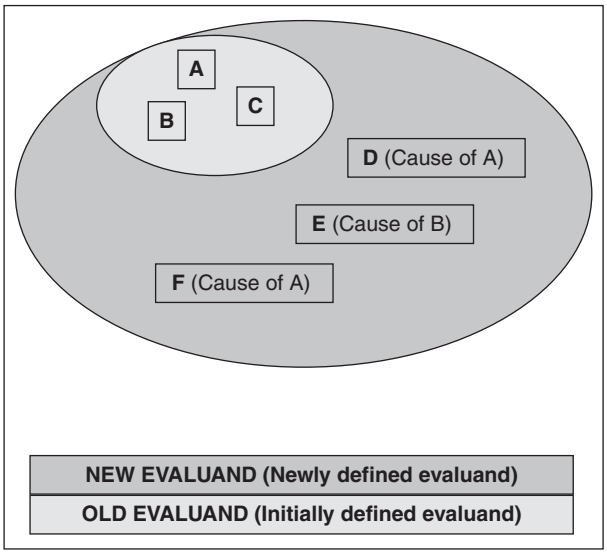

Figure 2. Improving our definition of the evaluand (and related set of criteria)

> reactivity with other chemical agents, reaction to heat, and so on. Social science concepts are no different. . . . Since one cannot avoid causal hypotheses when building concepts one must be as conscious as possible about them. (Goertz, 2006)

Criteria and Theory in the Evaluation of Organizations

Defining the evaluand in terms of interaction with its external environment, and paying attention to the mechanisms that explain its reactions in selecting the relevant criteria, are particularly appropriate in the evaluation of organizations. In this section two applications are presented: the first pertains to the Italian University Reform of the late 1990s and the second is taken from the auto-evaluation of KCCM, the Meetings Programming and Documentation Service of FAO.

Evaluating Italian Athenaeums

During the Italian University Reform of the late 1990s, the government decided to allocate a small but progressively increasing share of the funding (the so-called riequilibrium quota) to the institutions according to their performance on the following criteria: the ability to attract students, teaching and research. The indicator selected to assess merit on the teaching criterion was the average number of exams passed by students (corrected to take account of the different numbers of exams that different degrees require) (Osservatorio, 1998). In a meta-evaluation study (Befani, 2006), questions arose as to the validity of this indicator as the sole empirical evidence of teaching performance, particularly in the light of analyses conducted on the several possible strategies that institutions could adopt to make students pass more exams.

While all of these strategies lead to an increase in the number of exams passed by students, not all of them are equally desirable: strategies 1 and 2 increase concerns of equity while strategies 4 and 5 lower preparation quality. By unravelling the several mechanisms responsible for variations in the selected indicator, one realizes that teaching performance cannot be evaluated solely on the

Table 1. Institutional strategies potentially adoptable by Athenaeums in order to increase the number of exams passed by students

Strategy	Consequences
1. Attract only the best and most motivated students	*Access Restriction 1:* discrimination against average or subpar students; loss in terms of equality of opportunities.
2. Enrol only full-time students	*Access Restriction 2:* discrimination against those who cannot afford to study full-time; loss in terms of equality of opportunities.
3. Enhance the interaction between students and professors, hiring the latter for extra classes or clarifications, advice, etc. to students	Students enjoy an *increase* in their *learning opportunities* and an *improvement* in the *quality of teacher—student interaction,* which has additional positive effects on their *motivation*
4. Simplify the exam contents, cutting the hardest parts to learn	Learning is made easier not by improvements in teaching or as in the strategy above, but by a selection of exam contents, *distorting the appropriateness (quality and/or quantity-wise) of knowledge* or skills that students should have
5. Encourage the examiners to be less demanding and more indulgent in evaluating student preparation	Students are allowed to 'slack' and be poorly prepared, which has an additional negative impact on their motivation. The *quality of* their *preparation decreases*

exams-passed dimension. Clearly, when we look at teaching, we also want to ensure quality of preparation, student motivation and the ability to reach relatively wide audiences.

With the mechanisms responsible for variation in the indicator in mind, the evaluator redefines the concept of teaching performance, from which s/he extracts a new, more complete and more representative set of criteria. Had s/he missed this stage and trusted the previous indicator as the sole recipient of the 'value stock', s/he would have reached the wrong conclusions: for example that the system was in great shape while it was in fact in a terrible one. In Table2 we can see that a situation in which a 'major increase in the number of exams passed' takes place, can either be unacceptable, poor or excellent depending on the mechanisms that produce it.

Table 2. Criteria (Outcome + Mechanism) and standards of teaching performance

	Final outcome/result	Mechanisms producing increase in no. of exams that students pass
Excellent	Major increase	Major improvement in teaching or in quality of student—teacher interaction
Good	Considerable increase	Considerable improvement in quality of student— teacher interaction
Fair	Moderate increase	Moderate improvement in quality of student—teacher interaction
Poor	Major increase	Access restriction OR quality decrease
Unacceptable	Major increase	Access restriction AND quality decrease

Proceeding backward on the causal chain that leads to the passing of the exam and deciding which strategies leading to the desired outcome are good and what are not, allows the evaluator to judge as 'poor' something that s/he was previously judging as 'good' or 'excellent'.

Generalizing these ideas to performance evaluation in institutions, we can see that setting the product aside from the process as if they were two logically distinct entities, and failing to take into account either of the two while selecting evaluation criteria, puts us in danger of overlooking potentially significant dimensions in terms of value attribution. It is thus advisable to define the evaluand in terms of 'process leading to product' rather than process or product separately, and take into account both the performance level in the final outcome and the mechanisms that produce it.

The Evaluation of the Meetings Programming and Documentation Service (KCCM) of FAO

Similar reflections are inspired by the auto-evaluation of the Meetings Programming and Documentation Service of FAO (KCCM). In December 2007 KCCM launched an auto-evaluation aimed at measuring service performance during the previous three biennia (2002–7) and identifying areas for improvement. The process lasted about a year and a half and involved all the KCCM staff plus two external consultants in charge of defining the methodology and coordinating the process together with the Office of the Chief. A core team of 15 staff units – chosen in such a way as to represent the entire service staff – participated with the consultants at monthly official meetings during which virtually all aspects concerning the auto-evaluation were discussed: key questions to answer, evaluation criteria and standards, causal assumptions on service performance, data collection, names of interviewees, composition of focus groups, and many other methodological details. Drafts of the final report were regularly circulated among and discussed by the core team, in an attempt to enhance ownership of the auto-evaluation process among the service staff. The process was also designed in order to transfer evaluation skills from consultants to staff, which was a secondary goal of the auto-evaluation (Befani, 2009).

KCCM plans a variety of meetings, ensuring that they take place at the right dates so as not to overlap or interfere with other meetings, while guaranteeing the availability of a number of meeting services (rooms, technical services, translation and interpretation). Its main functions include Meeting Programming (e.g. compiling the meeting calendar and planning the supply of meeting services to be delivered at specific dates, taking into account the subplanning of logistic and linguistic services); the provision of interpretation on the day of the meeting; and translation services, both for in-session documents (to be translated in real time during the meeting) and pre-session ones (to be made available to participants and delegates a number of weeks before the meeting starts). KCCM also manages the translation of publications, official correspondence and administrative material. While meeting programming-related activities are funded by the regular budget, the delivery of linguistic services is ordered and paid for directly by FAO divisions via the so-called back-charge system. Fees – calculated on the basis of workload volume estimates and aimed at achieving a zero-balance between income and expenses at the end of the financial biennium – are established each biennium per each 1,000 words of translation or revision and per interpretation day.

The primary objectives of the auto-evaluation were twofold: on one hand, the service needed to find smarter ways of measuring performance than the basic quantitative indicators used for charging divisions and compiling financial reports (number of words translated, revised and interpretation days). On the other hand, the service needed to identify areas of improvement, with a preference for the 'quick-and-cheap' kind (given that the staff reduction trend could not be reversed and that – in the context of an FAO general reform under way – the long-term plans did not make much sense).

The first goal was thus searching for appropriate criteria to use in evaluating service performance. The service had no pre-established quantitative objectives to meet (other than the zero budget balance) and it was clear from the beginning that the number of interpretation days and the number of words translated did not capture the stakeholders' 'value stock' in its entirety. On the other hand, it was difficult to identify one single need those many activities were aimed at satisfying; the service had to respond in a satisfactory way to mainly five demands that divisions placed upon it in terms of their meeting organization-related needs: planning of meetings and of meeting logistical support; planning of translation services; planning of interpretation services; provision of translation services; and provision of interpretation services.

The first set of interviews (with both staff and users) was aimed at selecting a short list of criteria that were relevant to all or most of them; to find out what was relevant to both staff and users in evaluating the service. Two basic 'output' criteria, service quality and timeliness, were agreed on quite early in the process. Shortly afterwards, however, stakeholders started comparing current to past performance, and making assumptions on the factors influencing and explaining KCCM performance as they saw it. It was clear to them that evaluating the unit without taking into account the changes the rest of the organization had undergone in the last few years would not have made much sense. Internally, staff agreed with users that performance needed not only to be evaluated in terms of quality and timeliness, but it had to be explained and compared with reference to the different 'eras' in the history of the unit and the organization.

Translation of meeting documents, administrative material and publications. In some of the translation groups, translation jobs saw dramatic changes in the workflow from 2002 to 2007: while previously relying on a considerable number of internal staff translators, senior revisers were now faced with reduced staff and needed to outsource a high percentage of the workload to external freelance translators. While quality was previously ensured by the preparation and continued presence of staff at headquarters, outsourcing required that other quality-assurance mechanisms be put in place. Towards the end of the period under evaluation, the opportunity for revision had started to become the dividing line between a good translation and a poor one, while – as it became clear that the outsourcing trend was not slowing down in the near future – sustainability of the quality-assurance system began to be of concern. In particular, the age composition of the rosters (the extent to which they included experienced seniors along with an adequate number of junior professionals) was considered to be increasingly important in the medium–long term.

The definition of the evaluand (the translation function) varied with time and depended on the different types of workflow; it also varied within groups, as the changes were progressive and did not concern all the groups to the same extent or with the same speed. The evaluand could be split into two workflow models for which criteria took quite different importance weights: in the first, when the number of staff translators is high and outsourcing is low, revision of outsourced translation and roster sustainability are moderately important; on the other hand, as the number of staff translators is reduced and the outsourcing rate rises, revision and roster sustainability become extremely significant. These two criteria retain their importance even when the number of in-house translators stays high, if it is not high enough to avoid outsourcing. In Table 3 the final selection of relevant criteria for the translation of meeting documents, administrative material and publications is displayed, together with their importance levels according to the workflow model.

When comparing the groups in order to construct an overall rank, the different importance scores derived from knowledge on the quality-assurance mechanisms adopted in the different workflows influence the synthesizing algorithms. Translation Groups A, B and C report the grades shown in Table 4.

Table 3. Importance levels of criteria by workflow model

	Model 1: Internal Translation	Model 2: Outsourcing
1. Outsourcing (proportion of outsourced translation on total)	Low	High
2. Revision (proportion of outsourced translation that gets revised)	Moderately important	Extremely important
3. Staff translators (no. available)	High	—
4. Sustainability of the roster of freelance translators (average age of those employed)	Desirable	Very important

Table 4. Comparison between the groups

	Group A	Group B	Group C
1. Outsourcing	Good	Poor	Poor
2. Revision	Poor	Good	Good
3. In-house staff	Good	Good	Poor
4. Freelance Roster Sustainability	Poor	Poor	Good
Overall performance comparison	Group A the same as group C		Group B worse than Group C
Alternative comparison (based on no. of in-house staff)	Group A the same as Group B		Group C worse than both groups

All groups report two times good and two times poor, but their overall ranks differ. Group A – having a low outsourcing rate and a high number of staff translators – ensures quality through internal translation and belongs to model 1; therefore its 'poor' grades on revision and roster sustainability do not much affect its overall performance, as those two criteria have low importance in that model. Group B and Group C, on the other hand, adopt the recent kind of workflow, where quality needs to be ensured through revision and roster sustainability, and belong to model 2. Although they have a similar revision performance, Group B performs worse than Group C because it has a lower grade on one of the most important criteria (roster sustainability).

An evaluation of the translation function could thus not be executed properly without looking into the mechanisms activated to ensure quality assurance and the 'right mix of ingredients', or of 'what should go with what' in the two workflow models. If the evaluator had failed to acquire or take into account knowledge on the process, perhaps defining translation performance as the ability to translate in-house and thus taking the number of in-house translators as the sole criterion for translation performance, the evaluative conclusions would have been quite different. Groups A and B would have been considered equally good, while Group C would have been worse than both (see Table 4). Instead, the evaluand was redefined into two partially different entities, increasing the number of relevant criteria and differentiating the significance levels among them. Such redefinition was allowed by the acquisition of knowledge on the process leading to the outcome, in particular on which process led to the outcome in which conditions.

Table 5. Importance levels of criteria by context

	Context 1: Abundance of meeting officers	Context 2: Anyone can organize a meeting
1. Timeliness, reliability and completeness of request	Moderate—High	Low—Moderate
2. Extra work required to accommodate the request	Desirable	Extremely important
3. Knowledge of FAO facilities by request originator	High	Low
4. Satisfaction of request originator	Very important	Very important

Table 6. Comparison between the requests

	Request A	Request B	Request C	Request D
1. Timeliness, reliability and completeness of request	Fair	Fair	Excellent	Poor
2. Extra work required to accommodate the request	Poor	Good	Good	Unacceptable
3. Knowledge of FAO facilities by request originator	Poor	Good	Excellent	Poor
4. Satisfaction of request originator	Good	Good	Excellent	Poor
Overall performance comparison	Request A better than Request B		Request C the same as Request D	
Alternative performance comparison (using only user satisfaction)	Request A the same as Request B		Request C better than Request D	

Processing daily requests of meeting rooms and offices. A similar example concerns a function executed in the room-booking office: the 'processing of daily requests for meeting rooms, offices and assistance for internal meetings'. When the four criteria were identified, it became clear that not all of them had the same importance in different working conditions (see Table 5).

In the past, when staff rotation was lower and many divisions had meeting officer posts, the work conditions were simpler and easily predictable; in recent times, however, the office had to manage an increasing number of situations in which a considerable amount of information needed to be transferred to the division for every single request. The organization was moving towards an era in which meeting officers posts were being reduced and anyone, even an intern, was expected to be able to interact with KCCM in order to organize a meeting.

Drawing on knowledge about the process, the evaluator could distinguish between two categories of operating conditions, according to the different inputs coming from the request originators. The function was redefined into two slightly different evaluands, with different levels of importance for some of the evaluation criteria (see Table 5). When evaluating overall performance on different requests, the evaluator needed to synthesize the data according to the working conditions in which the request was handled, taking into account the different levels of importance for the criteria. Let us consider the four requests in Table 6.

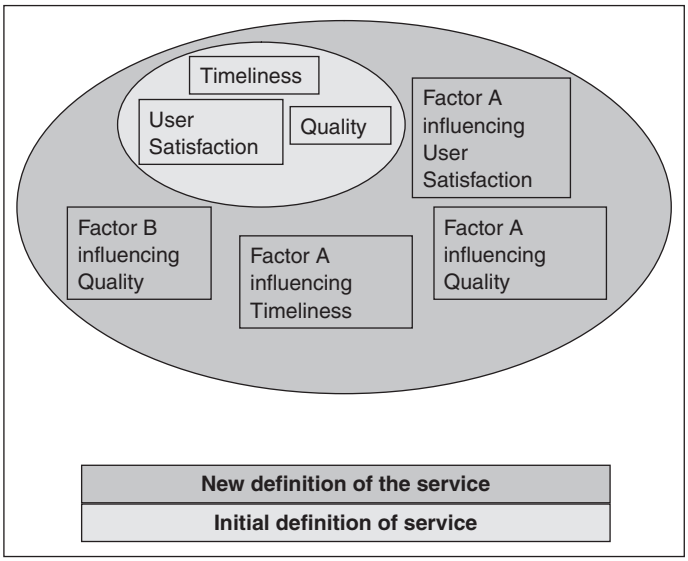

Figure 3. Improving our definition of the service (and related set of criteria)

While user satisfaction is good for both Requests A and B, and the timeliness and completeness of the request is also the same, the difference in knowledge on behalf of the originator made the service work harder for Request A than for Request B; performance on request A thus deserves to be evaluated better than that on Request B. In other words, Request A is handled in Context 2, where extra work is extremely important, while Request B is handled in Context 1, where extra work is only desirable. Request A has the best grades where it matters. Similarly, while user satisfaction is extremely different between Request C and Request D, one needs to take into account the different conditions in which the two requests were taking place that created more difficulties for Request D than for Request C; Request D was handled in Context 2 and needed an extreme amount of extra work, while Request C was handled in Context 1 and did not require the same effort. In other words, the bad grade on user satisfaction for Request D is compensated by the higher performance in terms of extra work.

Evaluating the service without taking these factors into account would have been short-sighted. Defining the service as a function aimed at achieving user satisfaction, using only the user satisfaction criterion, would have not reflected performance faithfully. The context represented by the knowledge and inputs of divisions plays a major part in the final outcome and cannot be overlooked. In brief, user satisfaction being equal, the service performed much better when it was able to accommodate requests placed in extreme, awkward situations, than when it executed simple tasks made easy by the knowledge and experience of request originators.

The function was thus redefined from a rigid structure that responds in an orderly and systematic way to user requests to a flexible, plastic entity that reacts differently to different inputs. As illustrated in Figure 3, the different contexts in which the unit operates were incorporated in the expanded list of evaluation criteria, and reflected in the different importance weights assigned to them. Failing to redefine the unit (the evaluand) would have resulted in different evaluative conclusions about its performance.

Conclusions

Through theoretical arguments and real-life applications, this article has shown that evaluation-specific logic and theory-based evaluation – far from being the opposite of one another – share the common goal of increasing the quality of evaluations, and there is no reason why they cannot or should not work together, creating synergies to benefit and improve the accuracy and validity of evaluative conclusions. While theory-based evaluation should draw from the evaluation-specific logic to ensure that value judgements are always made explicit and that evaluative research can always be clearly distinguished from 'ordinary' social research, the evaluation-specific logic should draw on theory to increase the chances that the most significant criteria are always correctly selected.

Note

1 It is ontological in the sense that it focuses on 'what constitutes a phenomenon'; causal because 'it identifies ontological attributes that play a key role in causal hypotheses, explanations, and mechanisms'. And realist because 'it involves an empirical analysis of the phenomenon' (Goertz, 2006).

References

Befani, B. (2006) 'Valutare l'Università: Un nuovo approccio orientate alla teoria: il caso italiano', unpublished doctoral dissertation.

Befani, B. (2009) 'KCCM Autoevaluation: Final Report', FAO Internal Circulation.

Bickman, L. (1987) 'The Functions of Program Theory', *New Directions for Program Evaluation* 33: 5–18.

Chen, Huey-Tsyh (1990) *Theory-Driven Evaluations*. London: SAGE.

Coleman, J. (1964) *Introduction to Mathematical Sociology*. Glencoe: Free Press.

Davidson, E. J. (2005) *Evaluation Methodology Basics: The Nuts and Bolts of Sound Evaluation*. London: SAGE.

Donaldson, S. I. (2007) *Program Theory-Driven Evaluation Science: Strategies and Applications*. New York: Lawrence Erlbaum.

Elster, J. (1998) 'A Plea for Mechanisms', in P. Hedström and R. Swedberg (eds) *Social Mechanisms: An Analytical Approach to Social Theory*, pp. 45–73. Cambridge: Cambridge University Press.

Goertz, G. (2006) *Social Science Concepts: A User's Guide*. Princeton: Princeton University Press.

Hedström, P. and R. Swedberg, eds (1998) *Social Mechanisms: An Analytical Approach to Social Theory*. Cambridge: Cambridge University Press.

Hedström, P. and R. Swedberg (1998) 'Social Mechanisms: An Introductory Essay', in P. Hedström and R. Swedberg (eds) *Social Mechanisms: An Analytical Approach to Social Theory*, pp. 1–31. Cambridge: Cambridge University Press.

Lipsey, M. W. (1993) 'Theory as a Method: Small Theories of Treatments', *New Directions for Program Evaluation* 57: 5–38.

Osservatorio per la valutazione del sistema universitario (1998) *Il riparto della quota di riequilibrio del FFO delle università: proposte per il triennio 1998–2000 doc 03/98.* Rome: Ministero dell'Università e della Ricerca Scientifica e Tecnologica (MURST).

Pawson, R. (2006) *Evidence-Based Policy: A Realist Perspective*. London: SAGE.

Pawson, R. and N. Tilley (1997) *Realistic Evaluation*. London: SAGE.

Scriven, M. (1959) 'The Logic of Criteria', *Journal of Philosophy* 56(22): 857–68.

Scriven, M. (1991) *Evaluation Thesaurus (*4th edn). London: SAGE.

Scriven, M. (1995) 'The Logic of Evaluation and Evaluation Practice', *New Directions for Program Evaluation* 68: 49–70.

Weiss, C. H. (1972) *Evaluation* (2nd edn). Upper Saddle River, NJ: Prentice Hall.
Weiss, C. H. (1997) 'Theory-Based Evaluation: Past, Present, and Future', *New Directions for Evaluation* 76: 41–55.
Wholey, J. S., ed. (1987) *Organizational Excellence: Stimulating Quality and Communicating Value*. Lexington, MA: Lexington Books.

Barbara Befani is an Evaluation Specialist and works as an independent consultant. She has a European PhD in socioeconomic and statistical studies and has collaborated with universities, the European Commission, International Organizations and numerous national and local administrations. Her research interests are theory-oriented approaches, the evaluation-specific logic and small-n methods. [email: befani@gmail.com]