How National and International Agencies Create Confusion about What Evidence is Needed for Evaluation

Jerome Winston and Arunaselam Rasappan

Jerome Winston, Director, *Centre for Development & Research in Evaluation, International* (Australia). jerome.winston@cedre.org.my

Arunaselam Rasappan PhD, Senior Advisor & Executive Director, *Center for Development & Research in Evaluation, International* (Malaysia). arasappan@cedre.org.my

1.0   Introduction

The purpose of this paper is to provide an updated report on exploratory evaluation research conducted at three evaluation R&D centres in Australia, Asia and Africa[1]. The research uses content analysis to answer the evaluation research questions in Table 1. The paper opens with a brief review of the history and practice of evaluation. This provides background to the development of multiple, and sometimes, conflicting definitions of evaluation terms. The word limit for conference papers restricts our discussion to terms in Table 2. Additional discussion is included in the conference presentation.

Table 1.   Exploratory Evaluation Research Questions

1. To what extent do national and international agencies provide multiple and some-times conflicting advice and requirements about what evidence may be included in a satisfactory evaluation?
2. Which, if any, evaluation terms or phrases have been defined in substantially different ways, by the same or different national and international agencies, without their explaining (a) the use of non-conventional definitions and/or (b) the expected benefits or disadvantages of adopting one particular definition rather than another?
3. In addition to long-standing debates about the extent to which *both* qualitative *and* quantitative data provide satisfactory "evidence for evaluation", what other debates and assumptions appear have influenced the publication of multiple and sometimes conflicting definitions and advice regarding what constitutes satisfactory "evidence for evaluation"[2]?

Table 1. Focus of evaluation research being conducted at 3 evaluation R&D centres.

2.0   Development of Evaluation Theory and Practice

Since the 1960s, there have been debates – particularly in the USA – about what constitutes satisfactory evidence for evaluation. In 1978, Patton discussed these debates in the first edition of *Utilization-Focused Evaluation*. When the lead author visited the U.S. General Accounting Office (GAO) during the 1980s, a senior GAO officer described long-standing disagreements between members of the Evaluation Research Society (ERS) and the Evaluation Network (ENet) about what constitutes satisfactory evidence for evaluation. The strength of these debates may be indicated by the fact that when an independent committee was funded to develop what have become, in effect, national (some might say, international) evaluation standards, the committee was – and still is – known as "The Joint Committee on Standards for *Educational* Evaluation" (emphasis added).

---

[1] The third evaluation R&D centre that is participating in this research is *CeDRE Africa*. Previous conference papers: Balakrishnan & Rasappan (2006); Winston & Bhagwandas (2006).

[2] A sub-theme of this conference is "collecting evidence for evaluations".

Table 2. Definitions of Selected Evaluation Terms

Evaluation Term	The Evaluation Centre	Other Sources including the *Evaluation Thesaurus*	OECD/DAC (2002)
Meta-evaluation	Evaluation of an evaluation.	[T]he evaluation of the first-level evaluation. (Scriven, 1991: 210) [T]he evaluation of evaluations—indirectly, the evaluation of [evaluation practitioners]…[3](*ibid*: 228) [A] method of evaluation research examining evaluation methodologies, procedures, data analysis techniques, interpretation of results, and the validity and reliability of conclusions. (Martin, 1982)	The term is used for evaluations designed to aggregate findings from a series of evaluations. It can also be used to denote the evaluation of an evaluation to judge its quality and/or assess the performance of the evaluators.
Meta-analysis	Not defined.	[Citing Glass: [A] *particular* approach to synthesizing quantitative studies on a common topic…[while for qualitative studies, there] is no special methodology... (emphasis in original; Scriven, 1991: 228) [S]ummarization of research findings to come to some general conclusions regarding effects or outcomes... (Martin, 1982)	Not defined
Summative Evaluation	Evaluation designed to present conclusions about the merit or worth…and recommendations about whether [the program] should be retained, altered, or eliminated.	[C]onducted *after* completion…(for ongoing programs, that means after stabilization) and *for* the benefit of some *external* audience…most often decisions…[to] export (generalize), increase site support, continue site support, continue with conditions, continue with modifications, discontinue (emphasis in original).[4] (Scriven, 1991:340)	A study conducted at the end of an intervention (or a phase of that intervention) to determine the extent to which anticipated outcomes were produced. Summative evaluation is intended to provide information about the worth of the program.[5]

---

[3] See also Exhibit 2.

[4] Scriven wrote: [Summative evaluation] should not be confused with outcome evaluation, which is simply an evaluation focused on outcomes rather than on process – which could be either formative or summative….It should also not be confused with global (holistic) evaluation—summative evaluation may be global *or* analytical. Where a summative evaluation is done of a program that has stabilized but is still running, the aim is the same: to report *on* it, not to report *to* it. (emphases in the original, p. 340)

[5] See below for information about *merit* and *worth*.

Table 2. (continued)

Evaluation Term	The Evaluation Centre	Other Sources including *Evaluation Thesaurus* (Scriven, 1991)	OECD/DAC (2002)
Formative Evaluation	Evaluation designed and used to improve…, especially when it is still being developed.	[T]ypically conducted *during* the development or improvement of a program…often more than once, for the in-house staff …*with the intent to improve* (emphasis in original).[6] (Scriven, 1991:168-69)	Evaluation intended to improve performance, most often conducted during the implementation phase of projects or programs.
Goal-Free Evaluation	Evaluation of outcomes …without knowledge of the purposes or goals. [from Joint Committee, 1994]	In the pure form…, the [evaluation team] is not told the [program's] purpose but [has] the purpose of finding out what the program is actually *doing* without being cued as to what it is *trying* to do.[7] (Scriven, 1991:180)	Not defined
Merit	excellence of an object as assessed by its intrinsic qualities or performance. [from Joint Committee, 1994]	[A]n entity's inherent, intrinsic, context-free value… Determining an entity's merit may take place whenever a number of experts are assembled. (Lincoln & Guba, 1979) Merit concerns a thing's excellence. (Sufflebeam, 1997)	Although *merit* and *worth* were not defined as separate terms, a note, which refers to these terms, was added to the definition of *effectiveness*: ***Effectiveness…***Note: Also used as an aggregate measure of (or judgment about) the merit or worth of an activity, i.e. the extent to which an intervention has attained, or is expected to attain, its major relevant objectives efficiently in a sustainable fashion and with a positive institutional development impact.
Worth	value of an object in relation to a purpose [from Joint Committee, 1994] system-related or extrinsic value…[various sources]	Worth is…contextually determined, place-bound value…Worth can only be determined by viewing the entity in operation or on site. (Lincoln & Guba, 1979) Worth concerns…cost-effectiveness in meeting clients' needs. (Stufflebeam, 1997)	

---

[6] "[O]ne of the most useful kinds…is 'early warning summative', that is, an evaluation which is essentially a summative evaluation of an early version of the evaluand under development." (Scriven, 1991:169) Note that when Scriven refers to a "program", he is referring to a planned intervention for change and not to a 'program' with the structure of a 'program budget'.

[7] Scriven wrote, "Merit is determined by relating program effects to the relevant *needs* of the *impacted* population, rather than to the goals of the program..." "[Goal-free evaluation] could equally well be called "needs-based evaluation"…by contrast with **goal-based** (or "manager-oriented") **evaluation** (emphasis in original).

2.1   1970s & Early 1980s in Australia

In contrast to the strength of debates in the USA about the acceptability of using one or another type of evidence for evaluation, when the lead author began to work actively with members of the growing community of evaluation practitioners in Australia – during the decade from 1975 to 1985 – such debates were largely absent. During those years, when debates about evaluation methodology did surface within the nascent Australian evaluation community, these debates were almost always 'imported' by graduates, academics and consultants who had studied in the USA, where their support had been sought for one or another point of view.[8]

The lead author of this paper recalls working actively at that time – in Queensland, Victoria, South Australia and New South Wales – with academics and human service providers whose teaching and practice promoted a broad approach to the design of evaluation[9]. This approach was also encouraged by the newsletters of the Australian Evaluation Network[10], the workshops organised by the Evaluation Training Network (Victoria)[11] and the National Evaluation Conferences, held in Melbourne starting in 1982, under the leadership of Dr Anona Armstrong.

In summary, during the 1970s and early 1980s, Australian evaluation practitioners, teachers and trainers largely adopted an eclectic approach to evaluation design – particularly when the evaluation was planned and implemented in social work, public health, hospital, vocational and community nursing, recreation, community, clinical and organisational psychology, occupational health and safety, agriculture extension and other human services. Evaluation practitioners selected evaluation designs and methods of data collection and analysis because they were expected to provide useful findings.

2.2   After the Introduction of Program Budgeting

Australian evaluation practitioners experienced considerable freedom during the 1970s and early 1980s regarding their choice of evaluation design. This condition changed as state and federal governments introduced program budgeting in Australia during the early-to-mid 1980s. These management "reforms" were similar to administrative reforms that had been introduced in Canada (c. 1975) and later in the UK, Malaysia and NZ. These reforms shared some common requirements including budgeting for, and evaluation of, 'programs'.

While public administrators in some government and international organisations were already familiar with developments in evaluation theory and practice[12], others had limited knowledge about, and experience with, using program evaluation to prepare budget proposals and submissions. This gave rise to requests for central agencies to provide advice and publications about how to plan, implement and use the results of evaluations to support new program proposals and submissions.

---

[8] However, Australia was not immune to these debates, which were more common at that time amongst *university-based* psychologists and sociologists who were not engaged in evaluation practice.

[9] A multi-disciplinary, eclectic approach to evaluation practice was encouraged by senior Australian academics including: Edna Chamberlain (head of the School of Social Work) at the University of Queensland; Marie Mune (head of the School of Social Studies) at the South Australian Institute of Technology; Elaine Martin (academic leader in Social Work) at Flinders University; and especially by Francis Donvovan (head of the School of Social Work) and her social work colleagues at Preston Institute of Technology (later amalgamated with RMIT University).

[10] Edited from the late 1970s to the early 1980s by Jerome Winston.

[11] Founded in Victoria by Jerome Winston and Colin Sharp in the early 1980s.

[12] Eg, departments of education had access to detailed advice about applying Action Research to the evaluation of technical education.

2.3   New Manuals and Guidelines on Evaluation

The number of new evaluation guidelines, manuals and handbooks grew in response to continuing requests for advice from central agencies and donors. Throughout the 1980s and 1990s and into the 2000s, governments and international agencies continued to publish new evaluation manuals, guidelines and handbooks which included definitions of evaluation terms. Some governments and agencies went through cycles of publication: definitions and explanations that appeared in earlier publications might be changed without explanation. International donors continued, as they had during the previous decades, to publish their own guidelines, some of which focused narrowly on evaluation, while others focused more on their catchall term, "M&E" – "monitoring-and-evaluation[13].

By the mid-to-late 1980s, evaluation practitioners in Australia and elsewhere who worked within, or for, national or international agencies were increasingly expected to become familiar with the latest published government or donor agency definitions and requirements. The authors of these definitions did not provide evaluation practitioners with a consistent view of evaluation. While some of the authors appear to have been familiar with long-established, conventional definitions and largely retained them, other authors provided definitions that were 'non-conventional' – usually without explanation. As a result of these innovations, new – and sometimes conflicting – definitions became a source of possible confusion regarding what evidence would be suitable for one or another type of evaluation.

3.0   Multiple Definitions of Evaluation Terms

Table 2 shows examples of variations in published definitions of evaluation terms. Definitions in the first two columns are accepted as being 'conventional' definitions. Two examples of a gap between long-established, internationally accepted definitions of evaluation terms, on the one hand, and 'non-conventional' definitions published by a government or international agency, on the other, are provided by variations in published definitions of the terms *meta-evaluation* and *summative evaluation*.

3.1   Meta-evaluation

The term *meta-evaluation* was coined by Michael Scriven. Scriven's explanation of the genesis of this term appears as Exhibit 1. A conventional definition of *meta-evaluation* is "evaluation of an evaluation". As conventional definitions of *meta-evaluation* had been readily available for decades (see Table 2), evaluation practitioners may be surprised to know that in 2002 the OECD published the following double-headed definition:

> ***Meta-evaluation*** The term is used for evaluations designed to aggregate findings from a series of evaluations. It can also be used to denote the evaluation of an evaluation to judge its quality and/or assess the performance of the evaluators. (OECD/DAC, 2002)[14]

By the 1980s, there was already international acceptance of Scriven's definitions of three terms that describe an evaluation: *formative*, *summative* and *meta-evaluation*. Inasmuch as none of these terms refers to any type of data analysis, statistical or otherwise, it is curious that, by the 2000s, some authors had provided definitions of *meta-evaluation* that implied (or stated explicitly) that *meta-evaluation* is a form of *meta-analysis*.

---

[13] We have often heard the term "M&E" used as if it were a single word, "em-an-ee", without clearly differentiating between monitoring and evaluation.

[14] As described in an appendix to this paper, an even more non-conventional definition of meta-evaluation appeared in a 2006 textbook that was released by one of the leading publishers of evaluation textbooks.

Exhibit 1.	Genesis of the Term *Meta-Evaluation*

Some years ago, Michael Scriven led an evaluation workshop in Melbourne, following his participation as a keynote speaker at an evaluation conference. During the workshop Scriven reminded participants that his first degree, from the University of Melbourne, was in philosophy. With a smile, Scriven recommended that all evaluation practitioners should consider following this example and earn a degree in philosophy, in order to be prepared to practice e*valu*ation. Scriven recounted that during his undergraduate studies he learned about a word that is used in linguistics: *metalanguage*. A layman's definition of *metalanguage* is a language that is used to talk about and describe another language. Drawing a parallel, Scriven told the workshop that he decided to call an evaluation that talked about and described another evaluation, a *meta-evaluation*. That is, the <u>meta</u> in *meta-evaluation* was not related to the term *meta-analysis* which refers to a method of data analysis. The origin of the <u>meta</u> in *meta-evaluation* drew on Scriven's knowledge of linguistics – in particular, his knowledge of the term <u>metalanguage</u>.

Recounted by Jerome Winston who attended the Scriven workshop.

3.2 Summative Evaluation

As shown in Table 2, definitions of *formative* evaluation from OECD/DAC, The Evaluation Centre (TEC) and *Evaluation Thesaurus* are sufficiently similar to suggest that the authors of the OECD definitions may have been familiar with conventional definitions of at least some evaluation terms. However, the OECD definition of *summative* evaluation differs from conventional definitions. It specifies three criteria for a *summative* evaluation:

- "conducted at the end of an intervention (or a phase of that intervention)"
- "determine[s] the extent to which anticipated outcomes were produced"
- "provide[s] information about the worth of the program".

Timing of Summative Evaluation

Scriven's definition of *summative* evaluation specifies **for whom** the evaluation is conducted – its **purpose** being to report to an **external** audience – rather than **when** it should be conducted. His definition places no time constraints on when a summative evaluation may be conducted, other than after an intervention has "stabilised". Likewise, the TEC definition places no constraints on when a *summative* evaluation is conducted.

By contrast, the OECD glossary specifies explicitly that a *summative evaluation* is "conducted **at the end** of an intervention (or a phase of that intervention)" [emphasis added]. This definition appears to presume that the word *summative* implies a "summing up" that may only occur at particular times during the life of an intervention. This assumption is not shared by either of the conventional definitions shown in Table 2.

Goal-Free Summative Evaluation

According to the OECD glossary, *summative* evaluations are **goal-focused**, as they "determine the extent to which anticipated outcomes were produced". Adopting this definition, it would be impossible to commission a **goal-free** *summative* evaluation. Conventional definitions allow this possibility.[15] The

---

[15] See Table 2 for a definition of *goal-free*. Note that Scriven understood that a goal-free evaluation would focus on how well needs were met, rather than on how well management's goals and objectives were met. Given Scriven's contributions to defining evaluation terms and promoting goal-free evaluation, it is not

OECD glossary provides no information regarding the decision to vary Scriven's definition and preclude the option of commissioning a **goal-free** *summative* evaluation.[16]

3.3 Ongoing Research

Our ongoing evaluation research is currently extending Table 2, to show (a) how other agencies have defined and used the terms *summative*, *formative* and *meta-evaluation* and (b) how national and international agencies have defined the terms *effectiveness*, *efficiency*, *program*, *output* and *outcome* or *impact*.

References

Balakrishnan, Suresh & Arunaselam Rasappan (2006). "Evaluation Practices in International donor-funded projects", Malaysian Evaluation Society 2nd Evaluation Conference, Kuala Lumpur.

GAO (1983). "The Evaluation Synthesis", General Accounting Office, Washington, D.C.

GAO (1987). "Drinking Age Laws: An Evaluation Synthesis of Their Impact on Highway Safety", General Accounting Office, Washing, D.C.

GAO (1990). "Prospective Evaluation Methods: The Prospective Evaluation Synthesis", General Accounting Office, Washington, D.C.

Joint Committee (1994). *The Program Evaluation Standards*, 2nd ed., Joint Committee on Standards for Educational Evaluation, Sage, California.

Lincoln, Yvonna S. & Egon G. Guba (1979). "The Distinction between Merit and Worth in Evaluation", ERIC, ED183574, www.eric.ed.gov.

Martin, Paula H. (1982). "Abstract: Meta-Analysis, Meta-Evaluation and Secondary Analysis", ERIC, ED228280, www.eric.ed.gov.

OECD/DAC (2002). "Glossary of Key Terms in Evaluation and Results Based Management", OECD, Paris, prepared by the Development Assistance Committee (DAC), Working Party on Aid Evaluation.

Patton, Michael Quinn (1978). *Utilization-Focused Evaluation*, Sage Publications, Beverly Hills.

Scriven, Michael (1991). *Evaluation Thesaurus (4th edition)*, Sage Publications, Newbury Park, California.

Stufflebeam, Daniel L. (1997). "Strategies for Institutionalizing Evaluation: Revisited", The Evaluation Center, www.wmich.edu/evalctr/pubs/ops/ops18.html.

TEC. "Glossary", The Evaluation Center, website last updated April 1, 2008, http://ec.wmich.edu/Glossary/glossaryList.htm.

Winston, Jerome & Ranjit Bhagwandas (2006). "KPIs: Do They *Really* Work as *Strategic* Tools for Evaluation and RBM?" Malaysian Evaluation Society 2nd Evaluation Conference, Kuala Lumpur.

---

surprising that conventional definitions allow for goal-free planning of both *formative* and *summative* evaluations.

[16] In practice, few evaluations are conducted in complete ignorance of an intervention's goals, objectives and anticipated results. A predominantly goal-free approach can be achieved by (a) treating the intervention's goals and objectives as management tools and evaluating their utility rather than accepting them as fixed targets to be achieved and (b) most importantly, seeking evidence of significant unplanned or unintended consequences associated with implementing the intervention.

Appendix. Evaluation Textbook with Non-Conventional Definitions

In 2006 a leading publisher of evaluation textbooks and reference books released a textbook[17] that provided the following non-conventional descriptions of *meta-evaluation* and *meta-analysis,* the first taken from the main text and the other two from an appendix:

> In the field of evaluation, [meta-evaluation is] a sub-discipline…[that] focuses on the meta-evaluation of programs. Meta-evaluations are syntheses of existing studies in a given area, and are intended to summarize what we know about, for example, head start programs.

> **meta-evaluation**   a synthesis of existing program evaluation studies in a given area, designed to summarize current knowledge about a particular type of program.

> **meta-analysis**   *the same as meta-evaluation*; a synthesis of existing program evaluation studies in a given area, designed to summarize current knowledge about a particular type of program. (emphasis added)

The 2006 textbook definition of meta-evaluation is similar to the OECD's <u>primary</u> definition of meta-evaluation (OECD/DAC, 2002). However, unlike the OECD glossary, the textbook provides no hint that the term *meta-evaluation* might refer to anything other than *meta-analysis*. The textbook states explicitly that *meta-analysis* is "the same as *meta-evaluation*" (emphasis added).

Neither the OECD glossary nor the textbook provides a definition of the evaluation term, *evaluation synthesis*, which may be used describe – using the textbook's words – "a synthesis of existing program evaluation studies…" (See GAO, 1983; 1987; 1990) Nor did either glossary include a conventional definition of *meta-analysis* (See Martin, 1982; Scriven, 1991).

---

[17] In this conference paper, we do not name the publisher, textbook title or the author with whom we anticipate continuing correspondence.