Escudero Escorza, T. (2003). From tests to current evaluative research. One century, the XXth, of intense development of evaluation in education. *RELIEVE, v. 9*, n. 1. http://www.uv.es/RELIEVE/v9n1/RELIEVEv9n1_1eng.htm

RELIEVE

Revista ELectrónica de Investigación y EValuación Educativa

This article in spanish (original Version)

FROM TESTS TO CURRENT EVALUATIVE RESEARCH. ONE CENTURY, THE 20TH, OF INTENSE DEVELOPMENT OF EVALUATION IN EDUCATION

(Desde los tests hasta la investigación evaluativa actual. Un siglo, el XX, de intenso desarrollo de la evaluación en educación)

by

Tomás Escudero Escorza
(tescuder@unizar.es)

Articles record
About the authors
HTML format

Ficha artículo
Sobre los autores
Formato HTML

Abstract

This article presents a critical review about historical development in the field of educational evaluation during the XXth century. The main theoretical proposals are commented

Keywords

Evaluation, Evaluation Research, Evaluation Methods; Formative Evaluation, Summative Evaluation, Testing, Program Evaluation

Resumen

Este artículo presenta una revisión crítica del desarrollo histórico que ha tenido el ámbito de la evaluación educativa durante todo el siglo XX. Se analizan los principales propuestas teóricas planteadas.

Descriptores

Evaluación, Investigación evaluativa, Métodos de Evaluación, Evaluación Formativa, Evaluación Sumativa, Test, Evaluación de Programas

Introduction

In any discipline, an investigation into its history tends to be a fundamental road to understand its conception, status, functions, environment, etc. This fact is especially evident in the case of evaluation because it is a discipline that has suffered deep conceptual and functional transformations throughout history and, mainly, throughout the 20th century, in which we principally concentrate our analyses. In this sense, the diachronic approach to the concept is essential.

We will carry out the analysis centering ourselves on three positions that we could brand as classics in recent literature on the topic and that we use indistinctly, although we don't have the pretense of offering a synthesis position, but rather of exact use of all them, since the three positions have an impact on the same moments and key movements.

A position, maybe the more used in our context (Mateo et al., 1993; Hernández, 1993), is that which Madaus, Scriven, Stufflebeam and

other authors offer that tends to establish six periods in their works, beginning its analysis in the 19th century (Stufflebeam and Shinkfield, 1987; Madaus et al., 1991). They speak to us of: a) period of reform (1800-1900), b) efficiency and "testing" period (1900-1930), c) Tyler period (1930-1945), d) innocence period (1946-1956), e) expansion period (1957-1972) and f) the professionalization period (from 1973) that connects with the current situation.

Other authors like Cabrera (1986) and Salvador (1992) cite three major stages, taking as a central reference point the figure of Tyler in the second quarter of the 20th century. The stages before Tyler are referred to as precedents or antecedents, Tyler's stage is referred to as the birth, and those which follow are considered development stages.

Guba and his collaborators, mainly Yvonna Lincoln, highlight different generations. We would currently be in the fourth (Guba and Lincoln, 1989) which according to them is based on the paradigmatic constructivist focus and in the needs of those stakeholders, as a base to determine the information that is needed. The first generation is that of measurement that arrives up until the first third of this century, the second is that of description and the third that of the judgement or valuation.

After the historical analysis, as a complement and a revision of synthesis, we offer a concise summary of the more relevant evaluative focuses of the different models and positions that, with greater or lesser force, come to mind when we try to delimit what is today evaluation research in education

1. Precedents: before «tests» and measurement

Since antiquity instructive procedures in which the teachers used implicit references have been created and used, without an explicit theory of evaluation, to value and, overall, to distinguish and select students. Dubois (1970) and Coffman (1971) cite the procedures that were used in imperial China more than three thousand years ago to select high officials. Other authors such as Sundbery (1977) speak of passages with reference to evaluation in the Bible, while Blanco (1994) refers to the exams of the Greek and Roman teachers. But according to McReynold (1975), the most important book of antiquity regarding evaluation is the Tetrabiblos that is attributed to Ptolomeo. Cicero and San Agustín also introduce evaluative concepts and positions in their writings.

It is during the Middle Ages that the exams are introduced into the university environment with a more formal character. It is necessary to remember the famous public oral exams in presence of a tribunal, although they were only administered to those individuals with previous permission from their professors, with which the possibility of failure was practically nonexistent. In the Renaissance there are continued uses of selective procedures and Huarte of San Juan, in his *Exam of geniuses for the sciences*, defends the observation as a basic position of the evaluation (Rodríguez et al., 1995).

In the 18th century, as the demand and the access to education increases, the necessity of verifying individual merits is accentuated and educational institutions embark on elaborating and introducing norms on the use of written exams (Gil, 1992).

In the 19th century, national systems of education are established and graduation diplomas appear following the passing of exams (exams of the State). According to Max Weber (Barbier, 1993), a system of exams of a specific preparation validation arises to satisfy the needs of a new hierarchical and bureaucratized society. In the United States, in 1845, Horace Mann begins to use the first evaluative techniques in the form of written tests. They extend to the schools of Boston and begin the road toward more objective and explicit references with relation to certain reading-writing skills. However, it still is not an evaluation sustained in a theoretical focus, but rather, something that responds to routine practices,

frequently based on not very reliable instruments.

At the end of the 19th century, in 1897, a work of J.M. Rice appears that is usually pointed out as the first evaluative research in education (Mateo et al., 1993). It discussed a comparative analysis in American schools about the value of the instruction in the study of spelling, using as criteria the marks obtained in tests.

2. Psychometric tests

In the previous context, at the end of the 19th century, a great interest for scientific measurement of human behaviors is awakened. This is something that is framed in the renovating movement of methodology of human sciences, when assuming the positivism of the physical-natural sciences. In this sense, evaluation receives the same influences as other pedagogic disciplines related with measurement processes, as experimental and differential pedagogy (Cabrera, 1986).

The evaluative activity will be conditioned in a decisive way by diverse factors that converge in this moment, such as:

a) The blossoming of the positivistic and empirical philosophical currents that supported observation, experimentation, data, and facts as sources of the true knowledge. The demand for scientific rigor as well as for objectivity in the measure of human behavior (Planchard, 1960) appears and written tests are promoted as a means of combating the subjectivity of oral exams (Ahman and Cook, 1967).

b) The influence of evolutionist theories and Darwin's works, Galton and Cattel, supporting the measurement of the characteristics of individuals and the differences among them.

c) The development of the statistical methods that favored decisively the metric orientation of the time (Nunnally, 1978).

d) The development of the industrial society that empowered the necessity to find some accreditation and selection mechanisms of students, according to their knowledge.

Consequently with this state of things, in this period between the end of the 19th century and beginning of the 20th, an intense evaluative activity known as “testing” is developed that is defined by the following characteristics:

. Measurement and evaluation were interchangeable terms. In the practice it was only referred to as measurement.

. The objective was to detect and establish individual differences, inside the pattern of trait and attribute that characterized the psychological development of the time (Fernández Archers, 1981), that is to say, the discovery of differential punctuations to determine the subject's relative position inside the reference group.

. The performance tests, synonym of educational evaluation, were developed to establish individual discriminations, forgetting in great measure the representativeness and consistency with educational objectives. In the words of Guba and Lincoln (1982), evaluation and measure had little relationship with the school curriculum. The tests showed results about the students, but not about the curriculums with which they had been educated.

Within the educational field, some instruments of that time are highlighted, such Ayres and Freeman’s writing scales, Hillegas’ writing, Buckingan’s spelling, Wood’s mathematics, Thorndike and McCall’s reading, and Wood and McCall’s arithmetic (Planchard, 1960; Ahman and Cook, 1967; Ebel, 1977).

However, it was in the psychological tests where the efforts had a larger impact, being most likely the work of Thorndike (1904) that had most influence during the beginning of the 20th century. In France the works of Alfred Binet stand out, later revised by Terman at the Stanford University, on tests of cognitive capacities. Now we speak of the Stanford-Binet, one of the most well-known tests in the history of the psychometry.

Years later, with the recruitment necessities in the First World War, Arthur Otis directs a team that builds collective tests of general intelligence (Alpha for readers-writers and Beta for illiterate) and inventories of personality (Phillips, 1974).

After the war, the psychological tests are put to the service of social ends. The decade between 1920 and 1930 marks the highest point in «testing» due to the development of a multitude of standardized tests to measure all kinds of school abilities with relating external and explicit objectives. They are based on procedures of intelligence measurement to use with large numbers of students.

These standardized applications are welcomed in educational environments and McCall (1920) proposes that the teachers construct their own objective tests, to not find themselves at the mercy of the proposals made exclusively by external specialists.

This movement was effective in parallel to the improvement process of psychological tests with the development of statistical and factorial analysis. The fervor for «testing» began to decline at the start of the 40's and there even began to arise some hypercritical movements with these practices.

Guba and Lincoln (1989) refer to this evaluation as the first generation that can rightfully be called the generation of measurement. The role of the evaluator used to be technical, as supplier of measurement instruments. According to these authors, this first generation still remains alive because texts and publications that use indissoluble evaluation and measure still exist (Gronlund, 1985).

3. The birth of true educational evaluation: The great "Tylerian" reform

Before the revolution promoted by Ralph W. Tyler arrived, an idependent current known as docimology started during the 1920's in France (Pieron, 1968 and 1969; Bonboir, 1972) that supposed a first approach to true educational evaluation. Mainly the split between that which is taught and the goals of the instruction was criticized. The evaluation was left in the hands of the completely personal interpretation of the teacher. As a solution the following was proposed: a) the elaboration of taxonomies to formulate objectives, b) diversification of information sources, exams, academic files, test retaking techniques, and tests, c) unification of correction criteria beginning with the agreement between the correctors of the tests and d) revision of value judgements by means of such procedures as double correction, or the means of different correctors. As can be seen, we are dealing with criteria of good and valid measurement, in some cases, even advanced.

Nevertheless, Tyler is traditionally considered the father of educational evaluation (Joint Committee, 1981), for being the first in giving it a methodical vision, going beyond behaviorism, a trend of the time, the mere psychological evaluation. Between 1932 and 1940, in his famous *Eight-Year Study of Secondary Education* for the Progressive Education Association, published two years later (Smith and Tyler, 1942), he outlines the necessity of a scientific evaluation that serves to perfect the quality of education. The synthesis work is published some years later (Tyler, 1950), explaining in a clear way his idea of curriculum, and integrating in it his systematic method of educational evaluation, as the process emerged to determine in what measure the previously established objectives have been reached (sees you also Tyler, 1967 and 1969).

The currículum comes defined by the four following questions:

a) What *objectives* are desired?

b) With what *activities* can they be reached?

c) How can these experiences be *organized* efficiently?

d) How can it be *proved* that the objectives are reached?

And the good precise evaluation of the following conditions:

a) Clear proposal of *objectives*.

b) Determination of the *situations* in those that should show the expected behaviors.

c) Election of *appropriate instruments* of evaluation.

d) *Interpretation* of the test results.

e) Determination of the *reliability* and *objectivity* of the measures.

This evaluation is no longer a simple measurement because it supposes a value judgement of collected information. It alludes to, without further development, the decision making regarding the success or failure of the curriculum according to students' results. This theme is one that important evaluators such as Cronbach and Sufflebeam would take up some years later.

For Tyler, the central reference in the evaluation is the preestablished objectives which should be carefully defined in behavioral terms (Mager, 1962), keeping in mind that they should mark the student's individual development, but inside a socialization process.

The object of the evaluative process is to determine the change in the students, but its function is wider than making explicit this change to the very students, parents and teachers; it is also a means of informing about the effectiveness of the educational program and also about the teacher's continuing education. According to Guba and Lincoln (1989), it refers to the *second generation* of evaluation. Unfortunately, this evaluative global vision was not sufficiently appreciated, neither exploited, for those that used its works (Bloom et al., 1975; Guba and Lincoln, 1982).

In spite of the above-mentioned issues and that the tylerianan reforms were not always applied immediately, Tyler's ideas were well received by the specialists in curricular development and by the teachers. Their outline was rational and it was supported by a clear technology, easy to understand and apply (Guba and Lincoln, 1982; House, 1989) and it fit perfectly in the rationality of the task's analysis that began to be used with success in military educational environments (Gagné, 1971). In Spain, the positions of Tyler extended with the General Law of Education of 1970.

After the Second World War, a period of expansion and optimism occurs that Stufflebeam and Shinkfield (1987) have not doubted to qualify as "social irresponsibility", due to the great consumer waste after a time of recession. It is the well-known stage as that of the *innocence* (Madaus et al., 1991). Many institutions and educational services of all types are extended, a large quantity of standardized tests are produced, there are advances in measurement technology and in the statistical principles of experimental design (Gulliksen, 1950; Lindquist, 1953; Walberg and Haertel, 1990) and the famous *taxonomies* of educational objectives appear (Bloom et al., 1956; Krathwohl et al., 1964). However, at this time, the contribution of evaluation to the improvement of education is scarce due to the lack of coherent plans of action. Much is written about evaluation, but with scarce influence in the improvement of the instructional work. The true development of the tylerianan proposals came later (Anklebone, 1962; Popham and Baker, 1970; Fernández de Castro, 1973).

Ralph W. Tyler died February 18th of 1994, having lived more than ninety years, and after seven decades of fruitful contributions and services to evaluation, research, and to education in general. Some months before, in April of 1993, Pamela Perfumo, a graduate student of Stanford University, interviewed Tyler with the purpose of knowing his thoughts about the current development of evaluation and of the controversial topics surrounding it. This interview, conveniently prepared, was presented April 16, 1993 in the AERA Conference in Atlanta. Horowitz (1995) analyzes the content and the meaning of the mentioned interview, highlighting, among other things, the following aspects of Tyler's thoughts at the end of his days:

a) Necessity to carefully analyze the purposes of the evaluation before beginning to evaluate. The current positions of multiple and alternateve evaluations should be adjusted to this principle.

b) The most important purpose in evaluation of the students is to guide their learning, this is, to help them to learn. A comprehensive evaluation of all the significant aspects of their performance is necessary; it is not enough to make sure that they regularly do their daily work.

c) The portfolio is a valuable evaluation instrument, but it depends on its content. In any event, it is necessary to be cautious with the preponderance of a single evaluation procedure, including the portfolio, for its inability of embracing the whole spectrum of evaluable aspects.

d) True evaluation should be idiosyncratic, appropriate for the student's peculiarities and the center of learning. In rigor, the comparison of centers is not possible.

e) Teachers should report to parents on their educational action with students. To do this it is necessary to interact with them in a more frequent and more informal way.

Half of a century after Tyler revolutionized the world of educational evaluation, one can observe the strength, coherence, and validity of his thoughts. As we have just seen, his basic, conveniently up-to-date, ideas are easily connected to the most current trends in educational evaluation.

4. The development of the sixties

The sixties would bring new airs to educational evaluation, among other things because people began to lend interest to Tyler's calls for attention, related with the effectiveness of the programs and the intrinsic value of evaluation for the improvement of education.

At that time a certain conflict arises between the American society and its educational system, mainly because Russia got ahead in the space program, after the launching of Sputnik for the USSR in 1957. A certain disenchantment appears with public schools and pressure grows for accountability (MacDonald, 1976; Stenhouse, 1984). In 1958 a new law of educational defense is promulgated that provides many programs and means to evaluate them. In 1964 the Elementary and Secondary Education Act (ESEA) is established by *the National Study Committee on Evaluation* and, creates a new evaluation not only of students, but to have an impact on programs and global educational practice (Mateo et al., 1993; Rodríguez et al., 1995).

To improve the situation and to recapture the scientific and educational hegemony, millions of dollars of public funds were dedicated to subsidize new educational programs and initiatives of American public schools' personnel guided to improve the quality of teaching. (Popham, 1983; Rutman and Mowbray, 1983; Weiss, 1983). This movement was also strengthened by the development of new technological means (audiovisual, computers...) and that of programmed teaching whose educational possibilities awoke interest in education professionals (Rosenthal, 1976).

In the same way that the proliferation of social programs in the previous decade had impelled the evaluation of programs in the social field, the sixties would be fruitful in demand for evaluation in the field of education. This new dynamic into which evaluation enters, though centered on the students as individuals that learn with the object of valuation being their performance, will vary in its functions, focus, and interpretation according to the type of decision being sought after.

To a great extent, this strong American evaluator impulse is due to the beforementioned approval of the *Elementary and Secondary Education Act* (ESEA) in 1965 (Berk, 1981; Rutman, 1984). With this law started the first significant program for educational organization in the federal environment

of the United States, and it was specified that each one of the projects carried out with federal economic support should be evaluated annually, in order to justify future grants.

Along with the disenchantment of public school, it is necessary to point out the economic recession that characterized the final years of the sixties, and, mainly, the decade of the seventies. This caused that general population, as taxpayers, and the legislators themselves to worry about the effectiveness and the yield of the money that was used in improving the school system. At the end of the sixties, and as a consequence to the above-mentioned, a new movement enters the scene, the era of *Accountability* (Popham, 1980 and 1983; Rutman and Mowbray, 1983) that is fundamentally associated with the teaching personnel's responsibility in the achievement of established educational objectives. In fact, in the year 1973, the legislation of many American states instituted the obligation of controlling the achievement of educational objectives and the adoption of corrective measures in negative cases (MacDonald, 1976; Wilson et al., 1978). It is comprehensible that, outlined this way, this movement of accountability and school responsibility, gave way to a wave of protests on the part of educational personnel.

Popham (1980) offers another dimension of school responsibility, when he refers to the school decentralization movement during the last years of the sixties and beginning of the seventies. Large school districts were divided into smaller geographical areas, and, consequently, with a greater direct civic control on what happened in the schools.

As a consequence of this focusing of influence, the phenomenon of educational evaluation was expended considerably. The direct subject of evaluation continued being the student, but also included all those factors that converge in the educational process (the educational program in a wide sense, teacher, means, contents, learning experiences, organization, etc.), as well as the educational product itself.

As a result of these new necessities of evaluation, a period of reflection and of theoretical essays with spirit of clarifying the multidimensionality of the evaluative process was initiated during this time. These theoretical reflections would decisively enrich the conceptual and methodological environment of evaluation that together with the tremendous expansion of program evaluation that occurred during these years, will give way to the emergence of the new modality of applied research that today we refer to as *evaluation research*.

As landmarks of the time, it is necessary to highlight two essays for their decisive influence: Cronbach's article (1963), *Course improvement through evaluation*, and that of Scriven (1967), *The methodology of evaluation*. The wealth of evaluative ideas exposed in these works forces us to refer to them briefly.

Regarding the analysis that Cronbach makes of the concept, functions and methodology of evaluation, we highlight the following suggestions:

a) *Associate the concept of evaluation with decision making*. The author distinguishes three types of educational decisions which the evaluation serves: a) about the improvement of the program and instruction, b) about the students (necessities and final merits) and c) about administrative regulation over the quality of the system, teachers, organization, etc. In this way, Cronbach opens the conceptual and functional field of educational evaluation far beyond the conceptual framework given by Tyler, although he follows his line of thought.

b) Evaluation that is used to *improve a program while it is being applied*, contributes more to the development of education than evaluation used to estimate the value of the product of an already concluded program.

c) Put in question the necessity that evaluative studies be comparative. Among the objections to this type of study, the author highlights the fact that frequently the differ-

ences among the average grades are lower in inter-groups than in intra-groups, as well as others concerning the technical difficulties that present comparative designs in the educational framework. Cronbach pleads for some absolute criteria of comparison, outstanding the necessity of an evaluation with reference to the criteria when defending the valuation with relation to some very well-defined objectives and not comparison with other groups.

d) Great scale studies are questioned, since the differences among their treatments can be very large and prevent the clear discernment of the results' causes. Defended are the more well-controlled analytic studies that can be used to compare alternative versions of a program.

e) Methodologically Cronbach proposes that evaluation should include: 1) process studies - facts that take place in the classroom-; 2) measure of performance and attitudes - changes observed in the students - and 3) follow-up studies, that is, the later path continued by the students that have participated in the program.

f) From this point of view, evaluation techniques cannot be limited to performance tests. Questionnaires, interviews, systematic and non-systematic observation, essays, according to the author, occupy an important place in evaluation, in contrast to the almost exclusive use that was made of tests like techniques of information collection.

If these reflections by Cronbach were shocking, they were not less than those in Scriven's essay (1967). His prolific terminological distinctions vastly enlarged the semantic field of evaluation, and at the same time clarified the evaluative chore. Below we highlight the most significant contributions:

a) Difinitively established is the difference between evaluation as a methodological activity, that which the author names the goal of the evaluation and the functions of the evaluation in a particular context. In this way evaluation as a methodological activity is essentially the same, whatever it may be that we are evaluating. The objective of evaluation is invariant, its main aim is the process with which we estimate the value of something that is evaluated, while the functions of the evaluation can be vastly varied. These functions are related with the use that is made of the collected information.

b) Scriven points out two different functions that evaluation can adopt: the formative and the summative. He proposes the term of *formative evaluation* to describe an evaluation of a program in progress with the objective of improving it. The term of *summative evaluation* is the process oriented to check the effectiveness of the program and to make decisions about its continuity.

c) Another important contribution of Scriven is the criticism of the emphasis that evaluation gives to the attainment of previously established objectives, because if the objectives lack value, one doesn't have any interest to know to what extent they have been achieved. The need for evaluation to include the evaluation of suitable objectives as well as determining the degreee to which these have been reached is emphasized (Scriven, 1973 and 1974).

d) Scriven makes a clear distinction between intrinsic evaluation and extrinsic evaluation, two different forms of valuing an element of teaching. In an intrinsic evaluation, the element is valued for what it is, while in extrinsic evaluation the element is valued for the effects that it causes in the students. This distinction is very important when considering the criteria to be utilized, because in intrinsic evaluation the criteria are not formulated in terms of operative objectives, while it is done in extrinsic evaluation.

e) Scriven adopts a position contrary to Cronbach, defending the comparative character that evaluation studies should present. Along with Cronbach he acknowledges the technical problems that comparative studies involve and the difficulty in explaining the

differences among programs. However, Scriven considers that evaluation, as opposed to the mere description, implies to produce a judgement about the superiority or inferiority of what is evaluated with regard to its competitors or alternatives.

These two commented contributions decisively influenced the community of evaluators, impacting not only studies in the line of evaluation research, to which is preferably referred, but also in evaluation orientated to the individual, in the evaluation line such as assessment (Mateo, 1986). We are before the *third generation* of evaluation that, according to Guba and Lincoln (1989), is characterized by introducing valuation, *judgement*, as an intrinsic content in evaluation. Now the evaluator doesn't only analyze and describe reality, he also assesses and judges it in relation to different criteria.

During the sixties many other contributions appear that continue drawing an outline of a new evaluative conception that will be finished developing and, mainly, extending in later decades. It is perceived that the conceptual nucleus of evaluation is the valuation of the change in the student as an effect of a systematic educational situation, some well formulated objectives being the best criteria to assess this change. Likewise, one begins to pay attention not only to the desired results, but also to the lateral or undesired effects, and even to results or long term effects (Cronbach, 1963; Glaser, 1963; Scriven, 1967; Stake, 1967).

There was criticism of the operativization of objectives (Eisner, 1967 and 1969; Atkin, 1968). Some criticism was directed at the structure of the underlying assessment. Another was about centering the assessment of learning in the most easily measurable products. Finally, other critics focus on the low attention given to the affective domain, with greater difficulty in operativization. In spite of this, Tyler's evaluative model would experience great improvement in these years, with works on the educational objectives that would continue and perfect the road undertaken in 1956 by Bloom and collaborators (Mager, 1962 and 1973; Lindvall, 1964; Krathwohl et al., 1964; Glaser, 1965; Popham, 1970; Bloom et al., 1971; Gagné 1971). Among other things new ideas appeared about the evaluation of interaction in the classroom and about its effects in the achievement of the students (Baker, 1969).

Stake (1967) proposed his evaluation model, *the countenance model*, that follows the line of Tyler. However, Stake's is more complete when considering the discrepancies among that which is observed and expected in the "antecedents" and "transactions", and when facilitating some bases to formulate a hypothesis about the causes and the shortcomings in the final results. In his successive proposals, Stake would begin distancing himself from his initial positions.

Metfessell and Michael (1967) also presented a model of evaluation of the effectiveness of an educational program in which, still following the basic pattern of Tyler, proposed the use of a comprehensive list of diverse criteria. The evaluators could keep these criteria in mind at the moment of evaluation and, consequently, not be centered merely in the intellectual knowledge reached by the students.

Suchman (1967) emphasized the idea that evaluation should be based on objective data that are analyzed with scientific methodology, clarifying that scientific research is preferably theoretical and, in exchange, evaluation research is always applied. His main purpose was to discover the effectiveness, success or failure, of a program when comparing it with the proposed objectives and, in this way, trace the lines of its possible redefinition. According to Suchman, this evaluation research should keep in mind: a) the nature of the addressee of the objective and that of the objective itself, b) the necessary time in which the proposed change is carried out, c) the knowledge of if the prospective results are dispersed or concentrated and d) the methods that must be used to reach the objectives. Suchman also defends the use of external evaluators to avoid all types of misrepre-

sentation by the teachers highly involved in the instructional processes.

The emphasis on the objectives and their measurement will also bring about need for a new orientation to evaluation, the denominated *evaluation of criterial reference*. The distinction introduced by Glaser (1963) among measurements referring to norms and criteria would have an echo at the end of the sixties, precisely as a result of the new demands that were outlined by educational evaluation. In this way, for example, when Hambleton (1985) studies the differences among tests referring to the criteria and tests referring to the norm he points out for the first ones, in addition to the well known objectives of describing the subject's performance and making decisions regarding whether or not a particular content is known, another objective similar to that of valuing the effectiveness of a program.

Since the end of the sixties, specialists have spoken decisively in favor of criterial evaluation, as soon as that is the evaluation type that gives real and descriptive information of the individual's or individuals' status regarding the foreseen teaching objectives, as well as the evaluation of that status for comparison with a standard or criteria of desirable realizations, being irrelevant to the contrast effect, namely the results obtained by other individuals or group of individuals (Popham, 1970 and 1983; Mager, 1973; Carreño, 1977; Gronlund, 1985).

In the evaluative practices of this decade of the sixties, two performance levels are observed. We can qualify one as *evaluation orientated toward individuals*, fundamentally students and teachers. The other level is that of *evaluation orientated to decision making on the "instrument" or "treatment" or educational "program."* This last level, also impelled by the evaluation of programs in the social environment, will be the basis for the consolidation in the educational field of program evaluation pro and of evaluation research.

5. From the seventies: The consolidation of evaluation research

If one could characterize the theoretical contributions that specialists offer us during the 1970's, it would be with the proliferation of all kinds of models of evaluation that flood the bibliographical market, *evaluation models* that express the author's own points of view who proposes them on *what it is* and *how* it should behave as an evaluative process. It deals with a time characterized by conceptual and methodological plurality. Guba and Lincoln (1982) speak to us of more than forty models proposed in these years, and Mateo (1986) refers to the *proliferation of models*. These will enrich the evaluative vocabulary considerably, however, we share Popham's idea (1980) that some are too complicated and others use quite confusing jargon.

Some authors like Guba and Lincoln (1982), Pérez (1983) and in some measure House (1989), tend to classify these models in two large groups, quantitative and qualitative, but we think along with Nevo (1983) and Cabrera (1986) that the situation is much richer in nuances.

It is certain that those *two tendencies* are observed today in evaluative proposals, and that some models can be representative of them, but different models, considered particularly, differ more by highlighting or emphasizing some of the components of the evaluative process and by the particular interpretation that they lend to this process. It is from this perspective, to our understanding, how the different models should be seen and be valued for their respective contributions in the conceptual and methodological fields (Worthen and Sanders, 1973; Stufflebeam and Shinkfield, 1987; Arnal et al., 1992; Scriven, 1994).

There are various authors (Lewy, 1976; Popham, 1980; Cronbach, 1982; Anderson and Ball, 1983; De la Orden, 1985) that consider the models not as exclusive, but rather as complementary, and that the study of them (at least those that have turned out to be more practical) will cause the evaluator to adopt a wider and more understanding vision of his work. We, in

some moment have dared to speak of modellic approaches, more than of models, since it is each evaluator that finishes building his own model in each evaluative research as a function of the work type and the circumstances (Escudero, 1993).

In this movement of evaluation model proposals, it is necessary to distinguish two stages with marked conceptual and methodological differences. In a *first stage*, the proposals followed the line exposed by Tyler in his position that has come to be called "*Achievement of Goals*." Besides those already mentioned by Stake and Metfessell and Michael that correspond to the last years of the sixties, in this stage the proposal of Hammond (1983) and the *Model of Discrepancy* of Provus (1971) stand out. For these authors the proposed objectives continue being the fundamental criteria of evaluation, but they emphasize the necessity to contribute data on the consistency or discrepancy between the designed guidelines and their execution in the reality of the classroom.

Other models consider the evaluation process at the service of the instances that should make decisions. Notable examples of them include: probably the most famous and utilized of all, the C.I.P.P. (context, input, process and product), proposed by Stufflebeam and collaborators (1971) and the C.S.E. (takes its initials from the University of California's Center for the Study of Evaluation) directed by Alkin (1969). The conceptual and methodological contribution of these models is positively assessed among the community of evaluators (Popham, 1980; Guba and Lincoln, 1982; House, 1989). These authors go beyond evaluation centered in final results, given that in their proposals they suppose different evaluation types, according to the necessities of the decisions which they serve.

A *second stage* in the proliferation of models is one represented by the concept of *alternative models* that, with different conceptions of evaluation and methodology, continue appearing in the second half of the seventies. Among those highlighted include Stake's *Responsive Evaluation* (1975 and 1976), to which Guba and Lincoln adhere to (1982), MacDonald's *Democratic Evaluation* (1976), Parlett and Hamilton's *Evaluation as Illumination* (1977) and Eisner's *Evaluation as Artistic Criticism* (1985).

In general terms, this second group of evaluative models emphasizes the role of the evaluation *audience* and the relationship of the evaluator with it. The high-priority evaluation audience in these models is not who should make the decisions, like in the models orientated to decision making, neither the one responsible for elaborating the curricula or objectives, like in the models of achievement of goals. The high-priority audience is the participants of the program themselves. The relationship between the evaluator and the audience, in the words of Guba and Lincoln (1982), should be "transactional and phenomenological". We are referring to models that advocate an ethnographic evaluation, it is from here that the methodology that is considered more appropriate is that used in social anthropology (Parlett and Hamilton, 1977; Guba and Lincoln, 1982; Pérez 1983).

This summary of models from the proliferation period is enough to approach to the wide theoretical and methodological conceptual array that today is related with evaluation. This explains that when Nevo (1983 and 1989) tries to carry out a conceptualization of evaluation, starting with the revision of the specialized literature, attending topics such as: What is evaluation? What functions does it have? What is the purpose of the evaluation?... a single answer is not found to these questions. It is easily comprehensible that the demands that evaluation suggests of programs of a part, and evaluation for making decisions regarding the individuals of another, drive a great variety of real evaluative outlines used by teachers, directors, inspectors, and public administrators. But it is also certain that below this diversity lie different theoretical and methodological conceptions about evaluation. Different conceptions have given way to an opening and conceptual plurality in the field of evaluation in

several senses (Goatherd, 1986). Next we highlight the most outstanding points of this conceptual plurality.

a) *Different evaluation concepts*. On one hand, we have the classic definition given by Tyler exists: *evaluation as the process of determining the consistency grade between the realizations and the previously established objectives*, to which the models orientated toward the realization of goals correspond. This definition contrasts with the wider one that is advocated by the models orientated to decision making: *evaluation as the process of determining, obtaining, and providing relevant information to judge alternative decisions*, defended by Alkin (1969), Stufflebeam et al. (1971), MacDonald (1976) and Cronbach (1982).

Moreover, Scriven's concept of evaluation (1967), being the process of estimating the value or the merit of something, is recaptured by Cronbach (1982), Guba and Lincoln (1982), and House (1989), with the objective of pointing out the differences that would involve value judgements in the event of estimating *merit* (it would be linked to intrinsic characteristics of what is being evaluated) or *value* (being linked to the use and application that it would have for a certain context).

b) *Different criteria*. It is deduced from the previously noted definitions that the criterion to use for evaluation of the information also changes. From the point of view of the achievement of goals, a good and operative definition of the objectives constitutes the fundamental criterion. From the perspective of the decisions and situated inside a political context, Stufflebeam and collaborators, Alkin and MacDonald even end up suggesting the non-evaluation of the information on the part of the evaluator, being the decision maker responsible of doing it.

The definitions of evaluation that accentuate the determination of "*merit*" as an objective of evaluation use standard criteria for those on which the experts or professionals agree. It deals with models related to accreditation and professional judgement (Popham, 1980).

The authors (Stake, 1975; Parlett and Hamilton, 1977; Guba and Lincoln, 1982; House, 1983) that accentuate the evaluation process in the service of determining the "value" more than the "merit" of the entity or object evaluated, advocate that the fundamental criterion of valuation be the contextual necessities in those that it is introduced. In this way, Guba and Lincoln (1982) relate the terms of the valorative comparison; on one hand, the characteristics of the evaluated object and, on the other, the necessities, expectations and values of the group to those that it affects or with those that the evaluated object is related.

c) *Plurality of evaluative processes* depending on the theoretical perception that is maintained over the evaluation. The cited evaluation models as well as others, too numerous to be in the bibliography, represent different proposals to drive an evaluation.

d) *Plurality of evaluation objects*. As Nevo says (1983 and 1989), there exist two important conclusions about evaluation that are obtained from the revision of the bibliography. On one hand, anything can be an evaluation object and should not be limited to students and teachers and, on the other, a clear identification of the evaluation object is an important part of any evaluation design.

e) *Opening*, generally recognized by all the authors, of the necessary information in an evaluative process to hold not only the desired results, but rather to the *possible effects* of an educational program, intended or not. Even Scriven (1973 and 1974) proposes an evaluation in which one doesn't have in mind sought after objectives, but values all the possible effects. Opening also regarding the collection of information, not only of the final product, but also of the educational process. And opening in the consideration of different results of *short* and *long* scope.

Lastly, opening not only in considering cognitive results, but also the affective ones (Anderson and Ball, 1983).

f) *Plurality in the functions* of evaluation in the educational field, withdrawing the proposal of Scriven between formative and summative evaluation, and adding others of sociopolitical and administrative type (Nevo, 1983).

g) Differences in the role played by the evaluator, which has come to be called *internal evaluation vs. external evaluation*. Nevertheless, a direct relationship between the evaluator and the different audiences of the evaluation is recognized by most of the authors (Nevo, 1983; Weiss, 1983; Rutman, 1984).

h) *Plurality of the audience* of the evaluation and, consequently, *plurality in the evaluation reports*. From informal narrative reports to very structured reports (Anderson and Ball, 1983).

i) *Methodological plurality*. The methodological questions arise from the dimension of evaluation as evaluation research that comes defined, in great measure, by methodological diversity.

The previous summary identifies the contributions made to evaluation in the 1970's and 1980's, a time period that has been named the *time of professionalization* (Stufflebeam and Skinkfield, 1987; Madaus et al., 1991; Hernández, 1993; Mateo et al., 1993). In addition to the countless models of the seventies, it was deepened in the theoretical and practical positions and consolidated evaluation as *evaluation research* in the term previously defined. In this context appear many new specialized magazines such as *Educational Evaluation and Policy Analysis, Studies in Evaluation, Evaluation Review, New Directions for Program Evaluation, Evaluation and Program Planning, Evaluation News*, etc. Scientific associations related to the development of evaluation are founded and universities are beginning to offer courses and programs in evaluation research, not only in graduate degrees and doctorate programs, but also in study plans for undergraduate degrees.

6. The fourth generation, according to Guba and Lincoln

At the end of the 1980's, after this whole before-described development, Guba and Lincoln (1989) offer an evaluating alternative that they call *fourth generation*, seeking to overcome that which, according to these authors, are deficiencies of the three previous generations, such as a manager point of view of the evaluation, a scarce attention to the pluralism of values, and an excessive attachment to the positivist paradigm. The alternative of Guba and Lincoln is called *responsive and constructivist*, integrating somehow the responsive focus proposed originally by Stake (1975), and the *postmodern* epistomology of constructivism (Russell and Willinsky, 1997). The demands, the concerns and the matters of the individuals involved or responsible (*stakeholders*) serve as the organizational focus of evaluation (as a base to determine what information is needed) which is carried out within the methodological positions of the constructivist paradigm.

The use of the demands, concerns, and matters of those involved is necessary, according to Guba and Lincoln, because:

a) They are *risk groups* with regard to evaluation and their problems should be adecuately contemplated, so that they are protected in the face of such a risk.

b) The results can be used *against* those involved in different senses, mainly if they are outside of the process.

c) They are potential users of the resulting evaluation information.

d) They can enlarge and improve the range of evaluation.

e) A positive interaction is produced among the different individuals involved.

These authors justify the paradigmatic change because:

a) Conventional methodology doesn't contemplate the necessity of identifying the demands, concerns, and matters of the individuals involved.

b) To carry out the above-mentioned, a discovery posture is needed more than verification, typical of positivism.

c) Contextual factors are not kept sufficiently in mind.

d) Means are not provided for case by case valuations.

e) The supposed neutrality of conventional methodology is of doubtful utility when value judgements are looked concerning a social object.

Leaving these premises, the evaluator is responsible for certain tasks that he/she will carry out sequentially or in parallel, building an orderly and systematic work process. The basic responsibilities of the evaluator of the fourth generation are as follows:

1) To identify all individuals involved with risk in the evaluation.

2) To bring out for each group involved their conceptions about what was evaluated and their demands and concerns about this matter.

3) To provide a context and a hermeneutic methodology in order to be able to keep in mind, understand, and criticize the different concepts, demands, and concerns.

4) To generate the maximum possible agreement about the said concepts, demands, and concerns.

5) To prepare an agenda for the negotiation of topics not in consensus.

6) To collect and to provide the necessary information for the negotiation.

7) To form and mediate a forum of involved individuals for the negotiation.

8) To develop and elaborate reports for each group of involved individuals on the different agreements and resolutions about their own interests and of those of other groups (Stake, 1986; Zeller, 1987).

9) Redraft the evaluation whenever there are pending matters of resolution.

The proposal of Guba and Lincoln (1989) extends quite a bit in the explanation of the nature and characteristics of the constructivist paradigm in opposition with those of the positivist.

When one speaks of the steps or phases of evaluation in this fourth generation, their proponents mention twelve steps or phases, with different subphases in each one of these. These steps are the following:

1) Establishment of a *contract* with a sponsor or client.

. Identification of the client or sponsor of the evaluation.

. Identification of the object of the evaluation.

. Purpose of the evaluation (Guba and Lincoln, 1982).

. Agreement with the client over the type of evaluation.

. Identification of the audiences.

. Brief description of the employed methodology.

. Guaranty of access to records and documents.

. Agreement to guarantee the confidentiality and anonimity to where it is possible.

. Description of the report type to elaborate.

. Listing of technical specifications.

2) *Organization* to redraft the research.

. Selection and training of the appraisal team.

. Attainment of facilities and access to the information (Lincoln and Guba, 1985).

3) Identification of the *audiences* (Guba and Lincoln, 1982).

. Agents.

. Beneficiaries.

. Victims.

4) Development of *conjunct constructs* within each group or audience (Glaser and Strauss, 1967; Glaser, 1978; Lincoln and Guba, 1985).

5) *Contrast and development* of the conjunct constructs of the audiences.

. Documents and records.

. Observation.

. Professional literature.

. Circles of other audiences.

. Ethical construct of the evaluator.

6) *Classification* of the demands, concerns, and resolved matters.

7) *Establishment of priorities* in the unresolved topics.

8) *Collection* of information.

9) Preparation of the *agenda* for negotiation.

10) Development of the *negotiation*.

11) *Reports* (Zeller, 1987; Licoln and Guba, 1988).

12) Recycling/*review*.

To judge the quality of the evaluation, we are offered three focuses called *parallel*, the linked to the *hermeneutic process* and that of *authenticity*.

The *parallel* criteria are named this way because they try to be parallel to the criteria of rigor used for many years inside the conventional paradigm. These criteria have been: internal and external validity, reliability and objectivity. However, the criteron should be in agreement with the fundamental paradigm (Morgan, 1983). In the case of the fourth generation, the criteria that are offered are those of *credibility*, *transfer*, *dependence*, and *confirmation* (Lincoln and Guba, 1986).

The *credibility* criteria are parallel to that of internal validity, so that the isomorphism idea between the findings and reality is replaced by the isomorphism among the realities built by the audiences and the reconstructions of the evaluator appointed to them. To achieve this, several techniques exist, among them the following are highlighted: a) prolonged compromise, b) persistent observation, c) contrast with colleagues, d) analysis of negative cases (Kidder, 1981), e) progressive subjectivity and f) control of the members. The *transfer* can be seen as parallel to the external validity, the *dependence* is parallel to the reliability criterion and the *confirmation* can be seen as parallel to the *objectivity*.

Another way to judge the quality of evaluation is through an analysis of the process itself, something that fits with the hermeneutic paradigm, through a dialectical process.

However, these two classes of criteria, although useful, are not completely satisfactory for Guba and Lincoln that also defend with more insistence the criteria that they call of *authenticity*, also of the constructivist basis. These criteria include the following: a) impartiality, justice, b) ontologic authenticity, c) educational authenticity, d) catalytic authenticity and e) tactical authenticity (Lincoln and Guba, 1986).

We can complete this analysis of the fourth generation with the characteristics with which Guba and Lincoln define evaluation:

a) Evaluation is a *sociopolitical* process.

b) Evaluation is a combined process of *collaboration*.

c) Evaluation is a *teaching/learning* process.

d) Evaluation is a *continuous* process, *recursive* and *highly divergent*.

e) Evaluation is an *emergent* process.

f) Evaluation is a process with *unpredictable* results.

g) Evaluation is a process that creates *reality*.

In this evaluation the characteristics of the evaluator are a result of the first three generations, namely that of the *technitian*, the *analyst*, and the *judge*. However, these should be expanded with skills in order to gather and interpret qualitative data (Patton, 1980). These skills include those of a historian and enlightener and those of a mediator of judegements which serve to create a more active role as an evaluator in the concrete socio-political concontext.

Russell and Willinsky (1997) defend the potentialities of the fourth generation's position to develop alternative formulations of evaluating practice among those individuals involved, increasing the probability that evaluation serves to improve school teaching. This requires, on the part of the faculty, the recognition of other positions, besides his own, the implication of all since the beginning of the process and, on the other hand, the development of more pragmatic approaches of the conceptualization of Guba and Lincoln, adapted to the different school realities.

7. The new impulse around Stufflebeam

To finish this analytic-historical journey from the first attempts of educational measurement to current evaluation research in education, we want to gather the recommendations that come to us more recently from one of the figures of this field in the second half of the 20th century. We are referring to Daniel L. Stufflebeam, proposer of the CIPP model (the most used) at the end of the sixties, president of the *Joint Committee on Standars for Educational Evaluation* from 1975 to 1988, and current director of the *Evaluation Center of Western Michigan University* (headquarters of the Joint Committee) and of CREATE (*Center for Research on Educational Accountability and Teacher Evaluation*), a center favored and financed by the Department of Education of the American Government.

Gathering these recommendations (Stufflebeam, 1994, 1998, 1999, 2000 and 2001), into those that have been integrating ideas of diverse notable evaluators, we don't offer just one of the lastest contributions to the current conception of evaluation research in education; we complete in good measure the vision of the current, rich and plural panorama, after analyzing the fourth generation of Guba and Lincoln.

Stufflebeam parts from the four principles of the *Joint Committee* (1981 and 1988), that is, from the idea that any good work of evaluative research should be: a) *useful*, that is, to provide timely information and to influence, b) *feasible*, this is, it should suppose a reasonable effort and should be politically viable, c) *appropriate, adequate, legitimate*, this is, ethical and just with those individuals involved, and d) *sure and precise* when offering information and judgements on the object of evaluation. Also, evaluation is seen as "transdisciplinary," because it is applicable to many different disciplines and many diverse objects (Scriven, 1994).

Stufflebeam invokes the responsibility of the evaluator that should act according to principles accepted by the society and to professionalism criteria, to form judgements regarding the quality and educational value of the evaluated object that should assist the involved individuals in the interpretation and use of its information and judgements. However, it is also

their duty, and their right, to be on the margin of the fight and the political responsibility for the decision-making process and eventual conclusion.

To evaluate education in a modern society, Stufflebeam (1994) suggests that some basic approaches of reference should be taken, such as the following:

. *Educational necessities*. It is necessary to ask oneself if the education provided covers the necessities of the students and their families in all areas in view of basic rights, in this case, inside a democratic society (Nowakowski et al., 1985).

. *Fairness, Equity*. It is necessary to ask oneself if the system is fair and equal when providing educational services, access, the achievement of goals, development of aspirations, and the coverage for all sectors of the community (Kellagan, 1982).

. *Feasibility*. It is necessary to question the efficiency of the use and distribution of resources, the adaptation and viability of the legal norms, the commitment and participation of those individuals involved, and everything that makes it possible for the educational effort to produce the maximum of possible fruits.

. E*xcellence as a permanently sought after objective*. The improvement of quality, starting from the analysis of the past and present practices is one of the foundations of the evaluation research.

Considering the reference point of these criteria and their derivations, Stufflebeam summarizes a series of recommendations to carry out good evaluative research and to improve the educational system. These recommendations consist of the following:

1) Evaluation plans should satisfy the four requirements of *utility*, *feasibility*, *legitimacy* and *precision* (Joint Committee, 1981 and 1988).

2) Educational entities should be examined for their integration and service to the *principles of democratic society*, equality, well-being, etc.

3) Educational entities should be valued in terms of their *merit* (intrinsic value, quality regarding general criteria) as much as their *value* (extrinsic value, quality and service for a particular context) (Guba and Lincoln, 1982; Scriven, 1991), as well as for their *significance* in the reality of the context in which it is located. Scriven (1998) points out that using other habitual denominations, merit has fairly good equivalence with the term quality, value with that of cost-effective relationship, and significance with that of importance. In any event, the three concepts depend on the context, specially the one refers to significance, meaning that understanding the difference between dependence on the context and arbitrariness is part of the understanding of the evaluation's logic.

4) Evaluation of teachers, educational institutions, programs, etc, should always be related to their duties, responsibilities, and professional or institutional obligations, etc. Maybe one of the challenges that educational systems should tackle is the clearest most precise definition of these duties and responsibilities. Without it, the evaluation is problematic, even in the formative field (Scriven, 1991*a*).

5) Evaluative studies should have the ability to value to what measure teachers and educational institutions are *responsible* and they *account* for the execution of their duties and professional obligations (Scriven, 1994).

6) Evaluative studies should provide *direction for improvement*, because it is not enough to simply form a judgement about the merit or the value of something.

7) Collecting the previous points, all evaluative study should have formative and summative components.

8) Professional self-evaluation should be encouraged, providing the educators with the skills needed and favoring positive attitudes toward it (Madaus et al., 1991).

9) *Evaluation of the context* (necessities, opportunities, problems in an area,…) should be used in a *prospective* way, to locate the goals and objectives and to define priorities. Also, evaluation of the context should be used *retrospectively* to adequately judge the value of the services and educational results, in connection with the necessities of the students (Madaus et al., 1991; Scriven, 1991).

10) *Evaluation of the inputs* should be used in a *prospective* way, to assure the use of an appropriate range of approaches according to the necessities and plans.

11) *Evaluation of the process* should be used in a *prospective* way to improve the work plan, but also in a *retrospective* way to judge to what extent the quality of the process determines the reason for why the results are at one level or another (Stufflebean and Shinkfield, 1987).

12) *Evaluation of the product* is the means of identifying the desired and not desired results in the participants or affected by the evaluated object. A *prospective* valuation of the results is needed to guide the process and to detect areas of need. A *retrospective* evaluation of the product is needed to be able to gauge as a whole the merit and value of the evaluated object (Scriven, 1991; Webster and Edwards, 1993; Webster et al., 1994).

13) Evaluative studies should base themselves on *communication* and the substantive and functional *inclusion* of stakeholders with the key questions, criteria, discoveries, and implications of the evaluation, as well as in the promotion of the acceptance and the use of their results (Chelimsky, 1998). Moreover, evaluative studies should be conceptualized and used systematically as part of the long-term process of educational improvement (Alkin et al., 1979; Joint Committee, 1988; Stronge and Helm, 1991; Keefe, 1994) and as grounds for action against social discriminations (Mertens. 1999). *Empowerment Evaluation* which Fetterman defends (1994), is a procedure, of democratic base, of involved individuals participating in the evaluated program, to promote their autonomy in the resolution of their problems. Weiss (1998) alerts us that participative evaluation increases the probability that the results of the evaluation are used, but also that it is conservative in its conception, because it is difficult to think that those responsible for an organization put in question its foundation and the system of power. Their interest is generally the change of small things.

14) Evaluative studies should employ *multiple perspectives*, *multiple measures of results*, and *quantitative* as well as *qualitative* methods to collect and analyze the information. The plurality and complexity of the educational phenomenon makes the use of multiple and multidimensional approaches in evaluative studies necessary (Scriven, 1991).

15) Evaluative studies should be evaluated, including *formative metaevaluations* to improve their quality and use as well as *summative metaevaluations* to help users in the interpretation of their findings and to provide suggestions for the improvement of future evaluations (Joint Committee, 1981 and 1988; Madaus et al., 1991; Scriven, 1991; Stufflebeam, 2001).

These fifteen recommendations provide essential elements for an approach of the evaluative studies that Stufflebeam calls *objectivist* and that is based on the ethical theory that moral kindness is objective and independent of personal or merely human feelings.

Without entering in debate concerning these final evaluations of Stufflebeam, or initiating a comparative analysis with other proposals, for example with those of Guba and Lincoln (1989), we find it to be evident that the conceptions of evaluation research are diverse, de-

pending on the epistemologic origin. However, there appear some clear and convincing common elements within all the perspectives such as *contextualization*, *service to society*, *methodological diversity*, *attention*, *respect and participation of those involved*, etc., as well as a greater *professionalization* of the evaluators and a wider *institutionalization* of the studies (Worthen and Sanders, 1991).

Stufflebeam (1998) recognizes the conflict of the positions of the *Joint Committee on Standards for Educational Evaluation* with those of the present trends in evaluation denominated postmodernist. Besides Guba and Lincoln, this conflict is represented additionally by other recognized evaluators such as Mabry, Stake and Walker, but he doesn't accept that reasons exist for attitudes of scepticism and frustration with the current evaluative practice, because many domains of approximation exist and the development of evaluation standards is perfectly compatible with the attention given to the diverse group of involved individuals and their values, social contexts and methods. Stufflebeam defends a larger collaboration in the improvement of evaluations, establishing standards in participative way, because he believes that the approach of positions is possible, with important contributions from all points of view.

Weiss (1998) also takes similar positions when she suggests that constructivist ideas should cause us to think more carefully when using the results of the evaluations, synthesizing them and establishing generalizations. However, she doubts that everything has to be interpreted in exclusively individual terms, as many common elements exist among people, programs and institutions.

8. To conclude: synthesis of model and methodological approaches of evaluation and Scriven's final perspective

After this analysis of the development of evaluation throughout the 20th Century, it seems opportune, as a synthesis and conclusion, to gather and emphasize those that are considered the main models, methodological positions, designs, perspectives and current visions. His analysis, in a compact manner, is a necessary complement for such a historic vision that, due to its lineality, runs the risk of offering an artificially divided disciplinary image.

During the seventies and the surrounding years, we have seen an appearance of evaluative proposals that traditionally have been called *models* (Castle and Gento, 1995) and in some cases *designs* (Arnal et al., 1992) of evaluation research. We know that several dozens of these proposals existed in the abovementioned decade, though were very concentrated in the time. In fact, the issue of those proposed models for evaluation seems to be a practically closed topic for nearly twenty years. New models or proposals no longer arise, except for some exceptions that we see later on.

In spite of that said, models, methods and designs in specialized literature, mainly looking for their agreement classification with diverse approaches, paradigmatic origin, purpose, methodology, etc. continued being discussed. Also in the classifications, not only in the models, exists diversity, which proves that, besides academic dynamism in the area of evaluation research, certain theoretical weakness still exists in this respect.

We have previously pointed out (Escudero, 1993) that we agree with Nevo (1983 and 1989) in the appreciation that many of the approaches to the conceptualization of evaluation (for example, the responsive model, the goal-free model, and the model of discrepancies, etc.) have been denominated unduly as models although none of them has the grade of complexity and globality that the previously-mentioned concept should carry. That which a classic text in evaluation (Worthen and Sanders, 1973) designates as «contemporary models of evaluation» (to the well-known positions of Tyler, Scriven, Stake, Provus, Stufflebeam, etc), Stake himself (1981) says that it would be better to call it "persuasions" while House (1983) refers to "metaphors".

Norris (1993) notes that the concept model is used with certain lightness when referring to conception, approach or even evaluation method. De Miguel (1989), on the other hand, thinks that many of the so-called models are only descriptions of processes or approaches to evaluation programs. Darling-Hammond et al. (1989) use the term "model" due to habit, but they indicate that they don't do it in the precise meaning of the term in social sciences, this is, basing it on a structure of supposed theory-based interrelations. Finally, we will say that the very author of the CIPP model only uses this denomination in a systematic way to refer to his own model (Stufflebeam and Shinkfield, 1987), using the terms approach, method, etc., when referring to the others. For us, perhaps the term *evaluative approaches* is the most appropriate, even if we continue speaking of models and designs simply due to academic tradition.

Our idea is that when conducting evaluative research, we still don't have a selected handful of well-based, defined, structured and complete models, from which to choose one in particular. However, we do indeed have distinct modellic approaches and ample theoretical and empirical support that allow the evaluator to respond in an appropriate manner to the different matters that the research process outlines, helping to configure a *global plan*, a *coherent flowchart*, and a «model» scientifically robust to carry out its evaluation (Escudero, 1993). Which are the necessary matters to address in this process of modellic construction? Leaning on in the contributions of different authors (Worthen and Sanders, 1973; Nevo, 1989; Kogan, 1989; Smith and Haver, 1990), they should address and define their answer while building a model of evaluation research in the following aspects:

1) Object of the evaluation research.

2) Purpose, objectives.

3) Audiences/participants/clientele.

4) High-priority or preferential emphasis/aspects.

5) Criteria of merit or value.

6) Information to collect.

7) Methods of information collection.

8) Analysis methods.

9) Agents of the process.

10) Sequenciation of the process.

11) Reports/utilization of results.

12) Limits of the evaluation.

13) Evaluation of evaluation research itself / metaevaluation.

To define these elements it is logically necessary to look for the support of the different modellic approaches, methods, procedures, etc., that evaluation research has developed, mainly in recent decades.

Returning to the denominated models of the seventies and to their classifications, we can gather some of those that appeared in the last decade in our academic field, based on different authors. In this way, for example, Arnal and others (1992) offer a classification of what they denominate *designs of evaluation research*, revising those of diverse authors (Patton, 1980; Guba and Lincoln, 1982; Pérez, 1983; Stufflebeam and Shinkfield, 1987). The classification is as follows:

Chart 1 - Types of designs of educational research

Perspective	*Patton (1980)*	*Guba and Lincoln (1982)*	*Pérez (1983)*	*Stufflebeam and Shinkfield (1987)*	*Creating authors*

Empirical-analytic	Objectives System analysis	Objectives	Objectives System analysis	Objectives Scientific method	Tyler (1950) Rivlin (1971) Rossi and et al (1979) Suchman (1967)
Liable to complementarity	CIPP Artistic criticism Adversary proceedings UTOS	CIPP Artistic criticism UTOS	CIPP Artistic criticism Cronbach (1982)	Adversary proceedings	Stufflebeam (1966) Eisner (1971) Wolf (1974)
Humanistic - interpretive	Responsive Illumination Goal-free	Responsive Goal-free	Responsive Illumination Democratic	Responsive Illumination Goal-free	Stake (1975) Parlett and Hamilton (1977) Scriven (1967) MacDonald (1976)

For their part, Castillo and Gento (1995) offer a classification of "methods of evaluation" within each one of the paradigmas that they call conductivist-efficientist, humanistic and holistic. The following is a synthesis of these classifications:

Chart 2 - Model behaviorist-efficientist

Method / author	*Evaluative purpose*	*Dominant paradigm*	*Content of evaluation*	*Role of the evaluator*
Achievement objectives Tyler (1940)	Measurement of achieved objectives	Quantitative	Results	External technician
CIPP Stufflebeam (1967)	Information for making decisions	Mixed	C (context) I (input) P (process) P (product)	External technician
Countenance Stake (1967)	Valuation of results and process	Mixed	Antecedents, transactions, results	External technician
CSE Alkin (1969)	Information for determination of decisions	Mixed	Centered in achievements of necessities	External technician
Educational planning Cronbach (1982)	Valuation of process and product	Mixed	U (evaluation units) T (treatment) O (operations)	External technician

Chart 3 – Humanistic model

Method / author	*Evaluative purpose*	*Dominant paradigm*	*Content of evaluation*	*Role of the evaluator*
Customer service Scriven (1973)	Analysis of the client's necessities	Mixed	All the effects of the program	External evaluator of necessities of the client
Opposition Owens (1973), Wolf (1974)	Opinions for consensus decision	Mixed	Any aspect of the program	External referee of the debate
Artistic criticism Eisner (1981)	Critical interpretation ofeducational actions	Qualitative	. Context . Emergent processes . Relations of processes . Impact on context	External stimulator of interpretations

Chart 4 – Holistic model

Method / author	*Evaluative purpose*	*Domi nante paradigm*	*Content of evaluation*	*Role of the evaluator*
Responsive Evaluation Stake (1976)	Valuation of answer to necessities of participants	Qualitative	Result of total debate on program	External stimulator of the interpretation for individuals involved
Holistic evaluation MacDonald (1976)	Educational interpretation for improvement	Qualitative	Elements that configure educational action	External stimulator of the interpretation for individuals implied
Evaluation as Illumination Parlett and Hamilton (1977)	Illumination and understanding of the program's components	Qualitative	System of teaching and means of learning	External stimulator of the interpretation for individuals involved

Scriven (1994) also offers a classification of the "previously-mentioned models," before introducing his transdisciplinary perspective which will be commented on later. This author identifies *six visions* or alternative approaches in the "explosive" phase of the models, in addition to others that he refers to as "exotic" that range from models of jurisprudence to expert models. Next we succinctly comment on these visions and the "models" that are attributed to them.

The strong decision-making vision (Vision A) provides the researching evaluator with the objective of reaching evaluative conclusions that help he/she that should make decisions. Those that support this approach worry if the program will reach its objectives, but they continue questioning if such objectives cover the necessities that they should cover. This position is maintained, although not made explicit by Ralph Tyler and is extensively elaborated in the CIPP model (Stufflebeam et al., 1971).

According to the Tylerian position, the decisions regarding a program should be based on the degree of coincidence between the objectives and the results. The degree of change in students, which is usually the pursued objective, is the evaluation criteria in this case.

Contrary to Tyler, Stufflebeam offers a wider perspective of the contents to be evaluated. The following are the four dimensions that identify his model, *context* (C) where the program takes place or the location of the institution, *inputs* (I) elements and initial resources, *process* (P) that is necessary to continue toward the goal and the *product* (P) that is obtained. Also, it is established that the fundamental objective of evaluation research is improvement, decision-making for the improvement of each of the four beforementioned dimensions.

Scriven (1994) tells us that Stufflebeam has continued developing his perspective since the development of the CIPP. However, one of his collaborators, Guba, took a different direction later on, just as we have seen when analyzing the fourth generation of evaluation (Guba and Lincoln, 1989).

The weak vision of decision-making (Vision B) provides the evaluator with relevant information for the making of decisions, but doesn't force him to produce critical or evaluative conclusions for the objectives of the programs. The most genuine theoretical representative is Marv Alkin (1969) that defines evaluation as a factual process of collection and generation of information at the service of the individual who makes the decisions, but it is this person that has to make the evaluative conclusions. This position is logically popular among those that think that true science shouldn't or cannot enter into questions of value judgements. Alkin's pattern is known as CSE (Center for the Study of Evaluation), outlining the following phases: valuation of the necessities and fixation of the problem, planning of the program, evaluation of the instrumentization, evaluation of progresses and evaluation of results.

The relativist vision (Vision C) also maintains the distance of the evaluative conclusions, but using the frame of the clients' values, without a judgement on the part of the evaluator about those values or some reference to others. This vision and the previous one have been the road that has allowed to many social scientists, integration without problems in the "car" of evaluative research. In fact, one of the most utilized texts of evaluation in the field of social sciences (Rossi and Freeman, 1993), utilizes this perspective.

Visions B and C are the positions of scientists connected to a free conception of scientific values. On the other hand, those that subscribe vision A come from a different paradigm, probably due to their academic connection with history, philosophy of education, compared education and educational administration.

Some years ago Alkin (1991) revised her positions from two decades ago, but continued without including the terms of merit, value, or worth. He finishes defining a System of Information for the Administration (Management Information System-MIS) for the use of the individual that makes decisions, but he doesn't offer valuations in this respect.

The simplest form of the *relativist vision* (Vision C) is the one developed in Malcolm Provus' "discrepancy model" of evaluation (1971). The discrepancies are the divergences with the sequence of projected tasks and the foreseen temporization. This model is closely related to program control in the conventional sense; it is a type of simulation of an evaluation.

The vision of the fertile, rich, complete description (Vision D) is that which understands evaluation like an ethnografic or journalistic task in which the evaluator reports on what he/she sees without trying to produce valorative statements or to infer evaluative conclusions, not even in the frame of the client's values as in the relativist vision. This vision has been defended by Robert Stake as well as by many British theorists. It is a kind of naturalistic version of vision B, having something of relativist flavor and sometimes appears to be a precursor of the vision of the fourth generation. It is based on observation, in the observable, more than in inference. Recently it has been denominated as a *vision of the solid, strong*

description, to avoid the *rich* term that seems more evaluative.

In his first stage, Stake is tylerian in regard to evaluative conception centered on the outlined objectives, proposing the countenance model (Stake, 1967), as a total image of the evaluation. This tour around the three components, *antecedents*, *transactions* and *results*, elaborates two matrices of data, one of *description* and another of *judgement*. In that of *description*, intentions are gathered from one side and the observations from the other and in the *judgement* matrix, the norms, which are approved and the judgements, which are believed to be appropriate, are collected.

During the mid-seventies, Stake moves away from the tylerian tradition of concern for the objectives and revises his evaluation method toward a position that he qualifies as "*responsive*" (Stake, 1975 and 1975a), assuming that the objectives of the program can be modified over time with the purpose of offering a complete and holistic vision of the program and to *respond* to the problems and real questions that are posed by those involved in the program. According to Stufflebeam and Shinkfield (1987), this model made Stake the leader of a new school of evaluation that demands a model that is pluralistic, flexible, interactive, holistic, subjective, and orientated to service. This model suggests "customer service" proposed by Scriven (1973), valuing their necessities and expectations.

In a graphic way, Stake (1975a) proposes the phases of the method through a comparison of the hours on a clock, putting the first one at twelve o'clock and continuing with the following phases in clockwise direction. These phases are the following: 1) Speak with the clients, those responsible, and audiences, 2) Scope of the program, 3) Panorama of activities, 4) Purposes and interests, 5) Questions and problems, 6) Data to investigate the problems, 7) Observers, judges and instruments, 8) Antecedents, transactions and results, 9) Development of topics, descriptions and case studies, 10) Validation (confirmation), 11) Outline for the audience and 12) Gathering of formal reports. The evaluator can also follow the phases in a counterclockwise direction or in any other order.

In the responsive method the evaluator must interview the participants to know their points of view and to look for the convergence of the diverse perspectives. The evaluator will interpret the opinions and differences in points of view (Stecher and Davis, 1990) and present a wide range of opinions or judgements, instead of presenting his/her personal conclusions.

The vision of social process (Vision E) that crystallized more than two decades ago around a group from Stanford University, directed by Lee J. Cronbach (1980), plays down the importance of the summative orientation of evaluation (external decisions about the programs and accountability), emphasizing the *understanding*, *planning* and *improvement* of social programs to those that it serves. Their positions were clearly established in ninety-five theses that have had an enormous diffusion between the evaluators and the users of the evaluations.

As for the contents of the evaluation, Cronbach (1983) proposes that the following elements are planned and controlled:

. *Units* (U) that are subjected to evaluation, individuals or participant groups.

. *Treatment* (T) of the evaluation.

. *Operations* (O) that the evaluator carries out for the collection and analysis of data, as well as for the elaboration of conclusions.

. *Context* in which the program and its evaluation takes place.

In one specific evaluative research, several units, treatments, and operations can be given, that is, several (uto), inside a (UTO) universe of acceptable situations.

Ernie House (1989), a theorists y practitioner of evaluation, quite independent of the latest trends in fashion, also marked the social connection of the programs, but he was distin-

guished mainly for his emphasis of the most ethical and argumentational dimensions of evaluation, perhaps motivated by the absence of these facets in Cronbach's approaches and his collaborators.

The constructivist vision of the fourth generation (Vision F) it is the last of these six visions that Scriven describes (1994), being maintained by Guba and Lincoln (1989) and continued by many American and British evaluators. We have already seen that this vision rejects an evaluation guided by the search for quality, merit, value, etc., and favors the idea that it is the result of the construction by individuals and the negotiation of groups. According to Scriven this means that all types of scientific knowledge are suspicious, debatable and non-objective. The same thing happens to all analytic work, including his philosophical analysis. Scriven notes that Guba himself has always been aware of the potential "self-contradictions" of his position.

From this revision made by Scriven, there are some evaluative positions traditionally gathered and delt with by analysts. In this way, for example, Schuman (1967) offers an evaluative design based on the *scientific method* or, at least, in some variation or adaptation of it. Owens (1973) and Wolf (1974 and 1975) propose an *opposition* method or discussion that through a program, cause the emergence of two groups of evaluators, partisans and adversaries, to distribute pertinent information for decision makers. Eisner (1971, 1975 and 1981) outlines the evaluation in terms similar to the process of artistic criticism.

Scriven himself (1967 and 1973) proposed years ago to base evaluation on *customer service* and not so much on the foreseen goals, given that the unforeseen achievements are frequently more important than those that figure in the planning of the program. Because of this, he tends to denominate his focus as *evaluation without goals*. The evaluator determines the value or merit of the program to inform the users; it is something similar to an informative middleman (Scriven, 1980).

Evaluation as illumination (Parlett and Hamilton, 1977) has a holistic, descriptive and interpretive approach, with the pretense of illumination on a complex range of questions that they are given in an interactive way (Fernández, 1991). MacDonald's *democratic evaluation* (1971 and 1976), also denominated holistic, supposes the collaborative participation of those individuals involved, and contrast of opinions of the participants is presumed to be the fundamental evaluative element.

Scriven (1994) critically analyzes the six visions and shows himself to be closest to vision A, *the strong vision on decision-making*, represented fundamentally by the CIPP model of Stufflebeam and his positions. He claims that it is the one that comes closest to the *common sense vision* which is the one that working evaluators use in their own programs, in the same way that doctors work with patients, making it the best thing possible, independently of the type and of the patient's general state. Scriven wants to extend this vision with a vision or model that he denominates *transdisciplinary* and that he qualifies as significantly different from the previously-mentioned vision A and radically different from the others.

In the *transdisciplinary perspective*, evaluation research has two components: the group of application fields of the evaluation and the content of the discipline itself. Something similar to what happens to disciplines as the statistic and the measurement. Difinitively, evaluation research is a discipline that includes its own contents as well as those of many other disciplines; its concern for analysis and improvement extends to many disciplines, making it transdisciplinary.

This vision is *objectivist* like vision A and it defends that the evaluator determines the merit or the value of the program, of the personnel or of the researched products. In such a sense, it should be established in an explicit way and defend the logic used in the inference of the evaluative conclusions, starting from the defi-

nitional and factual premises. Likewise, the argumentational fallacies of the doctrine free of values should also be pursued (Evaluation Thesaurus, 1991).

Secondly, the *transdisciplinary perspective* is centered on the consumer rather than the agent or intermediary. The perspective is not exclusive to the consumer, but it does consider the consumer first as a justification of the program. In addition, it considers common good a primacy of evaluation. Beginning here, valuable information is also produced for the agent who decides and can analyze the results of a particular program or institution in relation to its initial objectives. This position not only lends legitimacy to the researcher when generating evaluative conclusions, but also creates a necessity to perform such an analysis in the majority of cases.

It is also a *widespread vision*, though not exactly a general vision, that includes the generalization of concepts in the field of knowledge and practice. From this perspective, evaluation research is much more than the evaluation of programs, processes, and institutions and impacts in many other objects. In a more detailed way, this widespread vision means that:

a) The distinctive application fields of the discipline are the programs, personnel, achievements, products, projects, administration, and the metaevaluación of everything.

b) Evaluation research impacts in all types of disciplines and the resulting practices.

c) Evaluation research is conducted on levels, from practical to conceptual.

d) The different fields of evaluation research have many levels of interconnection and overlapping. The evaluation of programs, personal, centers, etc., have many aspects in common.

The fourth distinct element of the transdisciplinary perspective of evaluation is that it concentrates on a *technical vision*. Evaluation not only requires technical support from many other disciplines, but it includes its own unique methodology. To conduct a proper evaluation, it probably is not necessary to be a great specialist in auxilliary techniques and their processes of results synthesis, consequences, and their placement in the evaluation process as a whole; however, it is a necessity to have a thorough general understanding.

This *transdisciplinary perspective* of Scriven's evaluation research (1994), coincides in great measure with the positions that we have defended in other moments (Escudero, 1996). We don't have positions contrary to the other visions in the same measure that Scriven does and, in fact, we consider from a *pragmatic position* that all the visions have strong points and that in any event, they contribute something useful to the conceptual understanding and the development of evaluation research. However, we do think that this modern vision of Scriven is solid and coherent and broadly accepted at the present time.

A critique could be made of this position of Scriven concerning the excessive relative emphasis placed on client orientation, that is in the user in the strict sense. We think that this orientation should be integrated inside an orientation to individuals involved, where different types and different audiences exist and, of course, the users in the sense of Scriven is one of the most important, but it seems to us that evaluative research today has a *more plural high-priority orientation* than that which is defended by this author.

Bibliographic references

Ahman, S. J. y Cook, M. D. (1967). Evaluating Pupil Growth. Principles of Tests Measurement. Boston, Ma.: Allyn and Bacon

Alkin, M. (1969). Evaluation theory development. Evaluation Comment, 2, 1, 2-7.

Alkin, M. (1991). Evaluation theory development: II. En M. McLaughlin, y Phillips (Eds.), Evaluation and education at quarter century (pp.91-112). Chicago: NSSE/ University of Chicago .

Anderson, S. B. y Ball, S. (1983). The profession and practice of program evaluation. San Francisco, Ca: Jossey-Bass.

Arnal, J.; Del Rincon, D. y Latorre, A. (1992). Investigación educativa. Fundamentos y Metodología. Barcelona: Labor

Atkin, J. M. (1968). Behavioral Objectives in curriculum design: A cautionary note. The Science Teacher, 35, 27-30.

Baker, L. R. (1969). Curriculum evaluation. Review of Educational Research, 39, 339-358.

Barbier, J. M. (1993). La evaluación en los procesos de formación. Barcelona: Paidos.

Berk, A. R. (Ed.) (1981). Educational evaluation methodology: The state of the art. Baltimore: The Hopkins University Press.

Blanco, F. (1994). La evaluación en la educación secundaria. Salamanca: Amasú Ediciones.

Bloom, B. S., Engelhaut, M.D., Furst, E.J., Hill, W.H. y Krathwohl, D.R. (1956). Taxonomy of educational objectives Handbook I; Cognitive domain. New York: Davis, McKay.

Bloom, B. S., Hastings, J. Th. y Madaus, F. (1971). Handbook on formative and summative evaluation of student learning. New York: McGraw-Hill.

Bloom, B.S., Hastings, T. y Madaus G. (1975). Evaluación del aprendizaje. Buenos Aires: Troquel.

Bonboir, A. (1972). La Docimologie. Paris: PUF.

Cabrera, F. (1986). Proyecto docente sobre técnicas de medición y evaluación educativas. Barcelona: Universidad de Barcelona.

Carreño, H. F. (1977). Enfoques y principios teóricos de la evaluación. Mexico: Trillas.

Castillo, S. y Gento, S. (1995). Modelos de evaluación de programas educativos. En A.Medina y L. M. Willar (Coord.), Evaluación de programas educativos, centros y profesores (pp.25-69). Madrid: Editorial Universitas, S. A..

Chelimsky, E. (1998). The role of experience in formulating theories of evaluation practice. American Journal of Evaluation 19, 1, 35-55.

Coffman, W. E. (1971). Essay examinations. En R.L. Thorndike (Ed.) Educational Measurement. Washington , DC: American Council on Education.

Cronbach, L. J. (1963). Course improvement through evaluation. Teachers College Record, 64, 672-683.

Cronbach, L. J. (1982). Designing evaluations of educational and social programs. Chicago: Jossey-Bass.

Cronbach, L. J., Hambron, S.R., Dornbusch, S.M., Hess, R.D., Hornick, R.C., Phillips, D.C., Walker, D.F. y Weiner, S.S. (1980). Towards reform in program evaluation: Aims, methods and institutional arrangements. San Francisco: Jossey-Bass.

Cronbach, L. J. y Suppes, P. (1969). Research for tomorrow's schools: Disciplined inquiry for education. New York: MacMillan.

Darling-Hammond, L., Wise, AE, & Pease, SR (1989). Teacher evaluation in the organizational context: A review of the literature. En E. R. House (Ed.) New directions in educational evaluation (pp.203-253). London: The Falmer Press.

De la Orden, A. (1985). Investigación evaluativa. En Arturo De la Orden. (Ed.), Investigación educativa. Diccionario de Ciencias de la Educación (pp.133-137). Madrid: Anaya.

De Miguel, M. (1989). Modelos de investigación sobre organizaciones educativas. Revista de Investigación Educativa, 7, 13, 21-56.

Dubois, P. H. (1970). A History of Psychological Testing. Boston: Allyn Bacon.

Ebel, R. L. (1977). Fundamentos de la medición educacional. Buenos Aires: Guadalupe.

Eisner, E. W. (1967). Educational objectives: Help or hindrance?. The School Review, 75, 250-260.

Eisner, E. W. (1969, Instructional and expressive educational objectives: their formulation and use in curriculum. En J. Popham (Ed.), Instructional objectives (pp. 1-18). Chicago: AERA.

Eisner, E. (1971). Emerging models for educational evaluation. School Review, 2.

Eisner, E. W. (1975). The perceptive eye: Toward the reformation of educational evaluation. Stanford, Ca: Stanford Evaluation Consortion.

Eisner, E. W. (1981). The methodology of qualitative evaluation: the case of educational

connoisseurship and educational criticism. Stanford, Ca: Stanford University .

Eisner, W. E. (1985). The art of educational evaluation. London: The Falmer Press.

Escudero, T. (1993). Enfoques modélicos en la evaluación de la enseñanza universitaria, Actas de las III Jornadas Nacionales de Didáctica niversitaria «Evaluación y Desarrollo Profesi nal» (pp. 5-59). Las Palmas: Servicio de Publicaciones, Universidad de Las Palmas.

Escudero, T. (1996). Proyecto docente e investigador. Zaragoza: Universidad de Zaragoza.

Fernández Ballesteros, R. (1981). Perspectivas históricas de la evaluación conductual. En R. Fernández, y J.A.I Carrobles. (Ed.), Evaluación conductual. Madrid: Ediciones Pirámide.

Fernández de Castro, J. (1973). La enseñanza programada. Madrid: CSIC.

Fernández, J. (1991). La evaluación de la calidad docente. En A. Medina (Coord.), Teoría y métodos de evaluación. Madrid: Cincel.

Fetterman, D. M. (1994). Empowerment evaluation. Evaluation Practice, 15, 1, 1-15.

Gagne, R. M. (1971). Las condiciones del aprendizaje. Madrid: Aguilar.

Gil, E. (1992). El sistema educativo de la Compañía de Jesús. La «Ratio Studiorum». Madrid: UPCO.

Glaser, B. G. (1978). Theoretical sensitivity. Mill Valley, Ca: Sociology Press.

Glaser, B. G. y Strauss, A. L. (1967). The discovery of grounded theory. Chicago: Aldine.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: some questions. American Psychologists, 18, 519-521.

Glaser, R. (Dir.) (1965). Teaching machines and programmed learning. Washington: National Education Association.

Gronlund, N. E. (1985). Measurement and evaluation in teaching. New York: MacMillan,

Guba, G. E. y Lincoln, Y. S. (1982). Effective evaluation. San Francisco: Jossey Bass Publishers.

Guba, E. G. y Lincoln, Y. S. (1989). Fourth Generation Evaluation. Newbury Park, Ca.: Sage Publications

Gullicksen, H. (1950). Theory of mental tests. New York: Wiley

Hambleton, R. K. (1985). Criterion-referenced measurement.En Encyclopedia of Educational Research. New York: McMillan.

Hammond, R. L. (1983). Evaluation at the local level. En B. R. Worthen y J. R. Sanders, Educational Evaluation: Theory and Practice. Worthington, Ohio: Charles A. Jones Publishing Company.

Hernández, F. (1993). Proyecto docente e investigador. Murcia.

Horowitz, R. (1995). A 75-year legacy on assessment: Reflections from an interview with Ralph W. Tyler. The Journal of Educational Research, 89, 2, 68-75.

House, E. R. (1983). How we think about evaluation. En House, E. R., Philosophy of evaluation. San Francisco, Ca.: Jossey-Bass

House, E. R. (1989). Evaluating with validity. Newbury Park, Ca.: Sage.

Joint Committee on Standards for Educational Evaluation (1981, Standards for evaluations of educational programs, projects, and materials. New York.: McGraw-Hill.

Joint Committee on Standards for Educational Evaluation (1988). The personnel evaluation standards. Newbury Park, CA.: Sage.

Keefe, J. (1994). School evaluation using the CASE-IMS model and improvement process. Studies in Educational Evaluation, 20,1, 55-67.

Kellaghan, T. (1982). La evaluación educativa. Bogotá: Universidad Pontificia Javierana.

Kidder, L. H. (1981). Qualitative research and quasi-experimental frameworks. En M. B. Brewer y B. E. Collins, (Eds.), Scientific inquiry and the social sciences. San Francisco: Jossey-Bass.

Kogan, M. (Ed.) (1989). Evaluating higher education. London: Jessica Kingsley Publishers.

Krathwohl, R. D., Blomm, B.S. y Masia, B.B. (1964). Taxonomy of educational objectives. Handbook II: Affective domain. New Cork: Davis McKay.

Lewy, A. (Ed.) (1976). Manual de evaluación formativa del currículo. Bogotá : Voluntad/Unesco.

Lincoln, y. S. y Guba, E. G. (1985). Naturalistic inquiry. Beverly Hills, CA: Sage.

Lincoln, y. S. y Guba, E. G. (1986). But is it rigorous? Trustworthiness and authenticity in naturalistic evaluation. en D. D. Williams (Ed.), Naturalistic evaluation. San Francisco: Jossey-Bass.

Lincoln, y. S. y Guba, E. G. (1988). Criteria for assessing naturalistic inquiries as products. Paper presented at the American Educational Research Association, New Orleans, LA.

Lindquist, E. F. (1953). Design and Analysis of Experiments in Psychology and Education. Boston, Ma.: Houghton Mifflin Company.

Lindvall, C. M., (Ed.) (1964). Defining educational objectives. Pittsburg: University of Pittsburgh Press.

Macdonald, B. (1971). The evaluation of the Humanities Curriculum Project: a holistic approach. Theory into Practice, 10, 3, 163-169.

Macdonald, B. (1976). Evaluation and the control of education. En D. Tawney (Ed.), Curriculum evaluation today: trends and implications, 125-136. London: McMillan.

Madaus, G. F. y otros (1991). Evaluation Models. Viewpoints on Educational and Human Services Evaluation. Hingham , Mass.: Kluwer-Nijhoff Publishing.

Mager, R. F. (1962). Preparing instructional objective. Palo Alto, CA.: Fearon.

Mager, R. F. (1973). Análisis de metas. México: Trillas.

Mann, H., (1845). Boston Grammar and Writing Schools. Common School Journal, October, 7, 19.

Martínez de Toda, M. J. (1991). Metaevaluación de necesidades educativas: Hacia un sistema de normas. Tesis doctoral. Madrid: Universidad Complutense.

Mateo, J. (1986). Proyecto docente e investigador. Barcelona: Universidad de Barcelona.

Mateo, J. y otros (1993). La evaluación en el aula universitaria. Zaragoza: ICE-Universidad de Zaragoza.

Mccall, W. A. (1920). A new kind of school examination. Journal of Educational Research, January.

McReynold, P. (1975). Advances in Psychological Assessment, vol. III.San Francisco: Jossey-Bass.

Mertens, D. M. (1999). Inclusive evaluation: Implications of transformative theory for evaluation. American Journal of Evaluation, 20, 1, 1-14.

Metfessel, N. S. y Michael, W. B. (1967). A paradigm involving multiple criterion measures for the evaluation of effectiveness of school programs. Educational and Psychological Measurement, 27, 931-943.

Morgan, G. (1983). Beyond method. Beverly Hills, Ca: Sage.

Nevo, D. (1983). The conceptualization of educational evaluation: An analytical review of the literature. Review of Educational Research, 53, 1, 117-128.

Nevo, D. (1989). The conceptualization of educational evaluation: An analytical review of the literature. En E. R. House (Ed.), New directions in educational evaluation. London : The Falmer Press, 15-29.

Norris, N. (1993). Understanding educational evaluation. London : Kogan Page/CARE, School of Education , University of East Anglia .

Nowalowski, Jeri, Mary Anne Bunda, Russell Working (1985). A Handbook of Educational Variables. Boston: Kluwer-Nijhoff .

Nunnally, J. C. (1978). Psychometric Theory. New York: McGraw Hill.

Owens, T. R. (1973). Educational evaluation by adversary proceedings. En E.R. House (Comp.), School Evaluation: The Politics and Process. Berkeley: McCutchan.

Parlett, M. y Hamilton , D. (1977). Evaluation as illumination: A new approach to the study of innovative programmes. En Hamilton D. y otros (Eds.), Beyond the numbers game. London: MacMillan.

Patton, M. Q. (1980). Qualitative Evaluation methods. Beverly Hills, Ca.: Sage

Pérez, A. (1983). Modelos contemporáneos de evaluación. En J. Gimeno y A. Pérez, La en-

señanza: su teoría y su práctica (pp. 426-449). Madrid: Akal

Phillips, R. C. (1974). Evaluation in education. Columbus, Ohio: Merril.

Pieron, H. (1968). Vocabulaire de la psychologie. Paris: PUF.

Pieron, H. (1969). Examens et Docimologie. PUF, Paris.

Planchard, E. (1960). La investigación pedagógica. Madrid: Ediciones Fas.

Popham, W. J. (1970). Establishing instructional goals. Englewood Cliffs, N. J.: Prentice Hall.

Popham, W. J. (1980). Problemas y técnicas de la evaluación educativa. Madrid: Anaya.

Popham, W. J. (1983). Evaluación basada en criterios. Madrid: Magisterio Español, S. A..

Popham, W. y Baker, E. L. (1970). Systematic Instruction. Englewood Cliffs, NJ.: Prentice Hall.

Provus, M. (1971). Discrepancy evaluation. For educational program improvement and assessment. Berkeley, Ca.: McCutchan Publishing Co.

Rivlin, A. M. (1971). Systematic thinking for social action. Washington, D.C.: Brookings Institute.

Rodríguez, T. y otros (1995). Evaluación de los aprendizajes. Aula Abierta Monografías 25. viedo: ICE-Universidad de Oviedo.

Rosenthal, J. E. (1976). Evaluation history. En S. B. Anderson, y otros (Eds.), Encyclopedia of Educational Evaluation. San Francisco: Jossey Bass Publishers.

Rossi, P. H. y Freeman, H. (1993). Evaluation: A Systematic Approach. Beverly Hills , Ca.: Sage.

Rossi, P. H. y otros (1979). Evaluation: A systematic approach. Beverly Hills , Ca.: Sage.

Russell, N. y Willinsky, J. (1997). Fourth generation educational evaluation: The impact of a post-modern paradigm on school based evaluation. Studies in Educational Evaluation, 23, 3, 187-199.

Rutman, L. (Ed.) (1984). Evaluation research methods: A base guide. Beverly Hills, Ca.: Sage.

Rutman, L. y Mowbray, G. (1983). Understanding program evaluation. Beverly Hills, Ca.: Sage.

Salvador, L. (1992). Proyecto docente., Universidad de Cantabria.

Scriven, M. (1967). The methodology of evaluation. En Perspectives of Curriculum Evaluation, (pp. 39-83). AERA Monograph 1. Chicago : Rand McNally and Company.

Scriven, M. (1973). Goal-free evaluation. En E. R. House (Ed.), School evaluation: The politics and process,(pp. 319-328). Berkeley, CA.: McCutchan.

Scriven, M. (1974). Prose and cons about goal-free evaluation. Evaluation Comment, 3, 1-4.

Scriven, M. (1980). The logic of evaluation. Inverness , Ca.: Edgepress

Scriven, M. (1991, Evaluation Thesaurus. Newbury Park , Ca.: Sage

Scriven, M. (1991a). Duties of the teacher. Kalamazoo, Mi.: Center for Research on Educational Accountability and Teacher Evaluation.

Scriven, M. (1994). Evaluation as a discipline. Studies in Educational Evaluation, 20, 1, 147-166.

Scriven, M. (1998). Minimalist theory: The least theory that practice require. American Journal of Evaluation 19, 1, 57-70.

Smith, E. R. y Tyler, R. W. (1942). Appraising and recording student progress. New York: Harper & Row,.

Smith, N. L. y Haver, d. M. (1990). The applicability of selected evaluation models to evolving investigative designs. Studies in Educational Evaluation, 16, 3, 489-500.

Stacher, B. M. y Davis, W. A. (1990). How to Focus on Evaluation. Newbury Park, Ca.: Sage.

Stake, R. E. (1967). The countenance of educational evaluation. Teacher College Record, 68, 523-540.

Stake, R. E. (1975a). Program evaluation: particularly responsive evaluation. Occasional Paper, 5. University of Western Michigan.

Stake, R. E. (1975b). Evaluating the arts in education: A responsive approach. Ohio: Merril .

Stake, R. E. (1976). A theoretical stament of responsive evaluation. Studies in Educational Evaluation, 2 , 19-22.

Stake, R. E. (1981). Setting standars for educational evaluators. Evaluation News, 22, 148-152.

Stake, R. E. (1986). Quieting reform. Urbana: University of Illinois Press.

Stenhouse, L. (1984). Investigación y desarrollo del curriculum. Madrid: Morata.

Stronge, J. H. y Helm, V. M. (1991). Evaluating professional support personnel in educational settings. Newbury Park, Ca: Sage.

Stufflebeam, D. L. (1966). A depth study of the evaluation requeriment. Theory into Practice, 5, 3, 121-134.

Stufflebeam, D. L. (1994). Introduction: Recommendations for improving evaluations in U. S. public schools. Studies in Educational Evaluation, 20, 1, 3-21.

Stufflebeam, D. L. (1998). Conflicts between standards-based and postmodernist evaluations: Toward rapprochement. Journal of Personnel Evaluation in Education, 12, 3, 287-296.

Stufflebeam, D. L. (1999). Using profesional standards to legally and ethically release evaluation findings. Studies in Educational Evaluation, 25, 4, 325-334.

Stufflebeam, D. L. (2000). Guidelines for developing evaluation checklists. Consultado en www.wmich.edu/evalctr/checklists/ el 15 de D ciembre de 2002.

Stufflebeam, D. L. (2001). The metaevaluation imperative. American Journal of Evaluation, 22, 2, 183-209.

Stufflebeam, D. L., Foley, WJ, Gephart, WJ, Guba, EG, Hammond, RL, Merriman, HO & Provus, MM (1971). Educational Evaluation and Decision-making, Itasca, Illinois : F. E. Peacock Publishing.

Stufflebeam, D. L. y Shinkfield, A. J. (1987). Evaluación sistemática. Guía teórica y práctica. Barcelona: Paidos/MEC.

Suchman, E. A. (1967). Evaluative Research: Principles and Practice in Public Service and Social Action Programs. New York : Russell Sage Foundation.

Sunberg, N. D. (1977). Assessment of person. Englewood Cliffs, N.J.: Prentice Hall.

Taba, H. (1962). Curriculum development. Theory and practice. New York: Harcourt Brace .

Thorndike, E. L,. (1904). An Introduction to the Theory of Mental and Social Measurements. New York: Teacher College Press, Columbia University .

Tyler, R. W. (1950). Basic principles of curriculum and instruction. Chicago: University of Chicago Press .

Tyler, R. W. (1967). Changing concepts of educational evaluation. En R. E. Stack (Comp.), Perspectives of curriculum evaluation. AERA Monograph Series Curriculum Evaluation, 1. Chicago: Rand McNally,.

Tyler, R. W., (Ed.) (1969). Educational evaluation: New roles, new means. Chicago: University of Chicago Press .

Walberg, H. J. y Haertel, G. D. (Ed.) (1990). The International Encyclopedia of Educational Evaluation. Oxford: Pergamon Press.

Webster, W. J. y Edwards, M. E. (1993). An accountability system for school improvement, Paper presented at the annual meeting (April) of the AERA. Atlanta, GA.

Webster, W. J., Mendro, R.L. y Almaguer, T.O. (1994). Effectiveness indices: A «value added» approach to measuring school effect. Studies in Educational Evaluation, 20, 1, 113-137.

Weiss, C. H. (1983). Investigación evaluativa. Métodos para determinar la eficiencia de los programas de acción. México: Trillas,.

Weiss, C. H. (1998). Have we learned anything new about the use of evaluation?. American Journal of Evaluation, 19, 1, 21-33.

Wilson, A. R. J. (1978). La evaluación de los objetivos. En J. A. R. Wilson (Ed.), Fundamentos psicológicos del aprendizaje y la enseñanza (pp. 549-578). Madrid: Anaya..

Wolf, R. L. (1974). The citizen as jurist: A new model of educational evaluation. Citizen Action in Education, 4.

Wolf, R. L. (1975). Trial by jury: a new evaluation method. Phi Delta Kappa, 57, 185-187.

Worthen, B. R. y Sanders, J. R. (1973). Educational Evaluation: Theory and Practice. Worthington, Ohio: Charles a. Jones Publishing Company.

Worthen, B. R. y Sanders, J. R. (1991). The changing face of educational evaluation, Theory into Practice, XXX, 1, 3-12.

Zeller, N. C. (1987). A rethoric for naturalistic inquiry. Ph. D. dissertation, Indiana University.

ABOUT THE AUTHORS / SOBRE LOS AUTORES

Tomás Escudero Escorza (tescuder@unizar.es). Head of the department of Methods of Research and Diagnostics in Education at the University of Zaragoza. He is a member of the Comíssion for Technical Coordination of Quality Planning of Universities. He has authored a large number of works and publications in the evaluative education field, particularly institutional evaluation.

ARTICLE RECORD / FICHA DEL ARTÍCULO

Reference / Referencia	Escudero, Tomás (2003). Desde los tests hasta la investigación evaluativa actual. Un siglo, el XX, de intenso desarrollo de la evaluación en educación. *Revista ELectrónica de Investigación y EValuación Educativa (RELIEVE), v. 9*, n. 1. http://www.uv.es/RELIEVE/v9n1/RELIEVEv9n1_1.htm. Consultado en (*poner fecha*).
Title / Título	Desde los tests hasta la investigación evaluativa actual. Un siglo, el XX, de intenso desarrollo de la evaluación en educación. [*From tests to current evaluative research. One century, the XXth, of intense development of evaluation in education*]
Authors / Autores	Tomás Escudero Escorza
Translator / Traductora	Laura M. McLeod
Review / Revista	Revista ELectrónica de Investigación y EValuación Educativa (RELIEVE), *v. 9*, n. 1
ISSN	1134-4032
Publication date / Fecha de publicación	2003 (**Reception Date**: 2003 January 10; **Publication Date**: 2003 Febr. 27)
Abstract / Resumen	*This article presents a review and state of art about development in educational evaluation in the XXth century. The main theoretical proposals are commented.* Este artículo presenta una revisión crítica del desarrollo histórico que ha tenido el ámbito de la evaluación educativa durante todo el siglo XX. Se analizan los principales propuestas teóricas planteadas..
Keywords Descriptores	*Evaluation, Evaluation Research, Evaluation Methods; Formative Evaluation Summative Evaluation, Testing, Program Evaluation.* Evaluación, Investigación evaluativa, Métodos de Evaluación, Evaluación Formativa, Evaluación Sumativa, Test, Evaluación de Programas
Institution / Institución	Universidad de Zaragoza (España)
Publication site / Dirección	http://www.uv.es/RELIEVE
Language / Idioma	Español and english version (Title, abstract and keywords in english). *English version added March, 21th 2006*

Escudero Escorza, T. (2003). From tests to current evaluative research. One century, the XXth, of intense development of evaluation in education. *RELIEVE, v. 9*, n. 1. http://www.uv.es/RELIEVE/v9n1/RELIEVEv9n1_1eng.htm

Revista **EL**ectrónica de **I**nvestigación y **EV**aluación **E**ducativa
(**RELIEVE**)

[ISSN: 1134-4032]