SPECIAL ISSUE
Historical and Contemporary Perspectives on Educational Evaluation

education policy analysis archives

A peer-reviewed, independent,
open access, multilingual journal

Arizona State University

Volume 26 Number 46 April 16, 2018 ISSN 1068-2341

The Many Functions of Evaluation in Education

D. C. Phillips
Stanford University
United States

Citation: Phillips, D. C. (2018). The many functions of evaluation in education. *Education Policy Analysis Archives, 26*(46). http://dx.doi.org/10.14507/epaa.26.3811 This article is part of the Special Issue, *Historical and Contemporary Perspectives on Educational Evaluation: Dialogues with the International Academy of Education,* guest edited by Lorin W. Anderson, Maria de Ibarrola, and D. C. Phillips.

Abstract: This paper focuses upon the many functions that are served by evaluations, and by the process of conducting them. Many forms or types of evaluation have evolved to serve these different functions; and a brief account is given of some of the most common of these forms and the issues or controversies that they have engendered. The discussion opens, after a brief historical note, by describing the differing views of Lee Cronbach and Michael Scriven about whether an evaluator should aim to educate stakeholders about the program so that they can make informed decisions about it, or whether evaluators should go further and make a value judgement about it. The discussion then moves on to the importance of not overlooking the unintended effects of a program that is under study; and after presenting a list of functions that evaluations can have, the remainder of the discussion deals with the "pros" and "cons" of, and the differences between, formative and summative evaluations.

Journal website: http://epaa.asu.edu/ojs/
Facebook: /EPAAA
Twitter: @epaa_aape

Manuscript received: 1/15/2018
Revisions received: 2/18/2018
Accepted: 2/18/2018

Keywords: formative evaluation; summative evaluation; value judgements; decision-oriented research; conclusion-oriented research; unintended effects; functions of evaluation

Las múltiples funciones de la evaluación en educación
Resumen: Este artículo se enfoca en las múltiples funciones que cumple la evaluación, y en los procesos para llevarlas a cabo. Diversas formas y tipos de evaluación han evolucionado para servir a estas diferentes funciones; el artículo ofrece una breve reseña de las más comunes y de los tópicos o controversias que han engendrado. La discusión se inicia, después de una breve nota histórica, describiendo las diferentes posiciones de Lee Cronbach y Michael Scriven acerca de si el evaluador debe orientarse a educar a los interesados en el programa por evaluar, de manera que puedan tomar decisiones informadas acerca del mismo, o si los evaluadores deben ir más allá y ofrecer juicios de valor sobre el programa. A partir de ahí, la discusión se dirige a recalcar la importancia de tomar en cuenta los efectos no previstos del programa bajo estudio. Después de presentar la lista de funciones que las evaluaciones pueden tener, el resto de la discusión trata de los "pros" y los "contras" y de las diferencias entre la evaluación formativa y la sumativa.
Palabras-clave: evaluación formativa; evaluación sumativa; juicios de valor; investigación orientada a la toma de decisiones; investigación orientada a conclusiones; efectos no deseados; funciones de la evaluación

As múltiplas funções da avaliação na educação
Resumo: Este artigo concentra-se nas múltiplas funções desempenhadas pela avaliação e nos processos para realizá-las. Várias formas e tipos de avaliação evoluíram para atender a essas diferentes funções; O artigo oferece uma breve revisão dos mais comuns e dos tópicos ou controvérsias que geraram. A discussão começa, após uma breve nota histórica, descrevendo as diferentes posições de Lee Cronbach e Michael Scriven sobre se o avaliador deve ser orientado a educar os interessados no programa a ser avaliado, para que possam tomar decisões informadas sobre ele. , ou se os avaliadores devem ir mais longe e oferecer juízos de valor sobre o programa. A partir daí, a discussão visa enfatizar a importância de levar em conta os efeitos imprevistos do programa em estudo. Depois de apresentar a lista de funções que as avaliações podem ter, o resto da discussão lida com os "prós e contras" e as diferenças entre avaliação formativa e somativa.
Palavras-chave: avaliação formativa; avaliação sumativa; juízos de valor; pesquisa orientada para tomada de decisão; pesquisa orientada para conclusões; efeitos indesejados; funções de avaliação

The Many Functions of Evaluation in Education

> Those who forget the past are destined to relive it. (Santayana)

In this introduction to a collection of articles that had their origins in papers delivered to an enthusiastic audience of Mexican educational researchers and evaluators, it seems appropriate to issue a reminder that they are not alone in facing the daunting issues that arise when they labor to gather relevant information that can be useful in improving their national system of education. For there are colleagues around the world who have faced parallel challenges, and who have done (or are

doing) exemplary evaluations that should be sources of inspiration. Many of the following chapters will serve as illustrative examples, and they illustrate not only the diversity of functions of educational evaluations but also the variety of forms that these evaluations can take—for nowadays, the field is enormous, covering program evaluation, evaluation of educational personnel, assessment and testing of students, as well as evaluation of educationally relevant materials.

Thus, as a start, it is important to bear in mind that—given the century or so of inquiry and experimentation—there are already in existence many hard-won insights into the design and conduct of evaluations of various kinds. No doubt these pearls of wisdom can be infused into the Mexican (and hopefully many other) contexts with profit, if this has not happened already. It is this hopeful thought (reinforced by Santayana's famous insight quoted above) that undergirds much of this introduction.

A Very Selective Pre-History

There can be little doubt that evaluations of some kind have been carried out ever since the dawn of civilization—for evaluation is a part of effective governing. Emperors, pharaohs, and monarchs have wanted to know the effectiveness of their generals and admirals, the state of their food supplies, the progress of their building and road-construction programs, and the like. But scholarly interest in the field of evaluation—as a special (and broad) field with a focus upon improving evaluations—dates back only to sometime in the late 19thcentury. For example, in the USA arguably a tentative start of educational program evaluation was Joseph Rice's work documenting the deficiencies of American schools around the end of the 19th century; and even more important in the development of the field was the famous "eight year study" of progressive education carried out in the 1930s by a team led by Ralph Tyler—for it was not only the findings of this study that were of interest, but the evaluation process and methodology as well.

However, it was not until the decades of the 1960s and 1970s that the field of educational program evaluation clearly emerged as a semi-independent specialty within the broader educational research community (indeed, commonly the field of educational evaluation was called "evaluation research"). In those days evaluations of educational programs and interventions were becoming more common, and were being required by government bureaucrats and funding agencies—but at the same time there also was increasing concern that many evaluations were shoddy pieces of work, and that even good ones rarely had an impact. The perception was that many (if not most) evaluation reports were filed away and left to gather dust! In short, the process of evaluation had almost become a rather meaningless ritual—well summarized in these words from Shakespeare: "full of sound and fury, (but) signifying nothing." As late as 1981 a group at Stanford, led by the great educational psychologist and measurement genius Lee J. Cronbach (who in his younger days had been one of Tyler's assistants), wrote that "evaluation has vital work to do, yet its institutions and its ruling conceptions are inadequate….Moreover, evaluators become enthralled by technique" (Cronbach and Associates, 1980, p. 1). It was the widespread perception that program evaluation was not living up to its potential that prompted the emergence of a number of groups around the USA whose focus was the improvement of evaluation methodology. My personal experience was with the group at Stanford—called the "Stanford Evaluation Consortium" (SEC)—which thrived for almost a decade before slowly morphing from a research enterprise into a training program; some examples discussed below are drawn from my experiences in the SEC.

It also needs to be stressed that while program evaluation was following the trajectory outlined above, the field of educational testing and measurement also was generating enormous

interest and was home to numerous major technical breakthroughs (as will become abundantly clear in several of the following chapters).

But to return to program evaluation: The group at Stanford was not alone, and around this time there was an explosion of theories and so-called "models" of evaluation, and several productive groups for studying program evaluation sprang up in addition to the SEC (McLaughlin & Phillips, 1991). It became apparent early on that a fundamental difference of opinion existed between some of these groups.

Aiding Decision-Making (Cronbach) versus Making Value-Judgements (Scriven)

It was during the 1960s that Lee Cronbach first formulated the key ideas about the nature of educational research, and of educational program evaluation, that will be the main focus of this introductory chapter. Writing with Patrick Suppes, he distinguished between *decision-oriented* research and *conclusion-oriented* research; pure scientific research is aimed at firmly establishing the truth of conclusions; while applied work is often aimed at reaching decisions about what to do in practical contexts. Some educational research is conclusion-oriented, but much—including evaluation—is decision-oriented. In 1963, in what was to become a famous essay, this distinction was clearly in Cronbach's mind:

> we may define evaluation broadly as the *collection and use of information to make decisions about an educational program*…. many types of decisions are to be made, and many types of information are useful. It becomes immediately apparent that evaluation is a diversified activity…. (Cronbach, 1963, p. 672)

This decision-oriented account of the overall function of evaluation brought Cronbach (and later, his colleagues in the SEC) into loggerheads with the prominent philosopher of science, Michael Scriven, who was emerging as another central theorist of program evaluation. For Scriven was a supporter of conclusion-oriented evaluation. (As an aside, it is worth underlining the status of the men I have mentioned—Suppes, Cronbach, and Scriven—by noting that all three of them became presidents of the American Educational Research Association.) Scriven's account stressed that the very term "evaluation" was built around the word "value", and thus—he argued—the

> key sense of the term [evaluation] refers to the process of determining the merit, worth, or value of something….The evaluation process normally involves … some integration or synthesis of the results to achieve an overall evaluation or set of evaluations. (Scriven, 1991a, p. 139)

Thus there were (and possibly still are) two major, rival models of what an evaluation should have as its central focus: the aiding of decision-making, or the actual making of a judgment of the value of an educational program or intervention.

This is not the place to present a detailed account of the intellectual disputes between these two giants—disputes that were colored by their personal antipathy. Suffice it to say that in Cronbach's view, in a pluralistic democracy it was not appropriate for the evaluator (a non-elected person with no overall social mandate or authority) to take on the role of adjudicating between the interests and values of different groups—different stakeholder groups—by pronouncing that a program or social intervention had a particular high or low value. In a pluralistic democracy, the set of values that were acted upon in making decisions about educational and social interventions was determined not by an evaluator but by the political process playing itself out. In this scenario, the

evaluator's role was to *provide all stakeholders with relevant information* so that they could participate in this political process of decision-making in an enlightened and more empowered way. In response, of course, Scriven regarded this position as an evasive one that shirks the evaluator's responsibility.

Unintended Effects

In effect Scriven was focusing on making value-judgments about what the program being evaluated actually accomplished (if anything); that is, he was focusing upon the degree to which a program attained its stated or manifest goals, and what, if any, were its positive and negative unintended consequences. It was here, in the realm of unintended consequences, that Scriven made an important contribution.

Like many major social theorists, philosophers, and others—Karl Popper for example, and for that matter also Lee Cronbach—Scriven realized that interventions often failed to have their intended effect, but nevertheless their unintended effects could perhaps be of vital import. However, being unintended, these side effects might be hard to detect and easily could be overlooked. This phenomenon is well known in medical research and evaluation, but is also significant in education. Thus, seen in this light, the common approach in evaluation of educational programs—having the main instrument for data collection consist of an achievement test containing a number of test items based on the content of the course or program under evaluation—is quite deficient in detecting unintended consequences. Examples abound, along the lines of these simple cases: A curriculum designed to familiarize students with Shakespearian drama might (or might not) be successful, when measured by an end-of-semester test, however an unintended consequence—not revealed by the test—might be that many students are "turned off" to the "Bard of Avon" and become determined to avoid contact with any of his plays again; or mathematics and science programs might be devised that unintentionally (and unexpectedly) deter many young women and ethnic minority students from pursuing these subjects in the future.

Scriven devised so-called *goal-free evaluation* to deal with this phenomenon; and after his work, no evaluation could afford to ignore unintended side effects! Here I cannot resist describing my own use of the goal-free approach, long before I became familiar with Scriven's writings. When I was a young academic in Australia, from time to time I visited student teachers who were on practice-teaching assignments in high schools near to the university. It was common on such occasions for the student teacher to provide the university evaluator with a plan of the lesson that was about to be taught, with the goals clearly specified; at the end of the lesson the university expert would give a formative and summative evaluation. Quite often I would adopt the following approach—although usually it was not quite as harsh as my outline makes it out to be:

> I would tell the trainee that instead of focusing on his or her goals—on what the lesson was designed to achieve—my comments would focus on what the lesson *actually did do*! (The results, I would say, are more important than the intentions.) Thus, on one occasion, my evaluation went roughly as follows: "During the forty minutes of the lesson, three romances bloomed between pupils sitting near the back of the classroom (though thankfully none of these romances were actually consummated during this time); two boys perfected the art of flicking small balls of rolled-up paper at the necks of classmates sitting a row or two in front of them; the weekend football results were widely discussed; about a third of the class had no idea what the lesson was about, nor did they seem to care, while about two-thirds of the pupils left the room with the view that animals carry out respiration while plants do not (but instead carry-out photosynthesis). Since this was the main effect, it was

possibly the main aim of the lesson." After a pause I would add: "What a pity this view about respiration is seriously mistaken!"

The Many Functions of Evaluations

Evaluations are themselves interventions—sometimes large-scale and expensive ones—and they often if not always take place in a social context that is marked by discord. There often are dedicated supporters of the program or intervention that is being evaluated, and there can be diehard opponents, who have ideological or other reasons firing their distaste for it; there are those who wish to see the program closed down so that the funds can be diverted to support some other cause that they deem more worthy; there may be stakeholder groups which see their own economic or political interests being bound up (positively or negatively) with the fate of the program; and researchers and policy analysts might be interested in the program for what it can tell them about the prospects for achieving social reform and increasing social justice.

Many chapters in this present volume illustrate—and enlarge upon—these points. The contributions by Servaas van der Berg and Kadriye Ercikan discuss evaluations that can yield data and policy insights that will be of value in the struggle to improve educational prospects for children in developing countries. (These articles too—as well as others—are worth close study for their clear and informative modes of reporting policy-relevant data.) The article by Maria de Ibarrola is a fascinating account of a highly discordant and politicized context in which researchers and evaluators had to work, a context in which it is understating matters to say that the different stakeholder groups had markedly different attitudes towards the program that was at stake. David Berliner discusses the dilemmas that need to be confronted and the choices that need to be made when the focus is on evaluation of teachers; and Lorin Anderson discusses the most widespread mode of evaluation of students—namely, the practice of assigning grades—and in a remarkably comprehensive discussion he throws light on the numerous factors that influence the assignment of grades and the many purposes that are hoped to be achieved by the comparison of student grades. Richard Shavelson, furthermore, reminds readers of the important (but often overlooked) truth that the choice of research methods to use, and the choice of features that need to be incorporated into the design of an evaluation, all depend upon the purpose of the study. And there is a further major complexity, again touched on by Shavelson: The design, choice of measuring instruments and materials, and even the selection of those who are working on the evaluation, must all take into account the fact that in most if not in all modern societies there is enormous cultural and political diversity—along with which there are enormous challenges. (For example, consider the difficulties associated with ensuring that Instructional materials and measuring instruments are culturally appropriate.) Sylvia Schmelkes illuminates matters such as these in her important article.

Michael Scriven was well aware of the general point being made here, namely that evaluations can have many functions in addition to assessing whether program goals had been attained, but he seemed to downplay a tad the discordant environment in which programs and their evaluations often were situated. Nevertheless, in discussing the many types of evaluation he brilliantly wrote of the similarities and differences between product evaluation, personnel evaluation, cost-effectiveness evaluation, component evaluation, dimension evaluation, global evaluation, among others (see Scriven, 1991a, 1991b). Throughout, he maintained his position regarding the importance of making of a value-judgement about a program in light of what it achieved or what it failed to achieve (whether the effect was intended or not).

Cronbach and his associates in the SEC also emphasized that evaluations could have many functions. Consider this passage:

> Evaluations are initiated for many purposes, sometimes conflicting ones: choosing a best prospect among several proposed lines of action, fine-tuning a program already in operation, maintaining quality control, forcing subordinates to comply with instructions, documenting that one's agency deserves its budget, creating support for a pet project, casting suspicion on a policy favored by political opponents…. (Cronbach and Associates, 1980, p. 13)

To which can be added: achieving a major delay in making a decision about terminating a program (until the new evaluation is complete), providing a mechanism wherein heated disputes among stakeholder groups could possibly cool down, gathering evidence for or against the firing of a senior administrator or program official, establishing credit or blame when a new administrator or political clique has taken over the program, determining if the program is actually being delivered in the manner intended by its designers.

The article by William Schubert, with an exhaustive bibliography, fills in a lacuna here; for he engagingly reminds us that evaluations take place in an intellectual context—which on occasion can be almost as discordant as the social and political one! Evaluators who are enamored of different background theories, paradigms or philosophies, research ideologies, and the like, are liable to see the program (or in his case, the curriculum) differently and therefore are also likely to ask different evaluative questions about it, and ultimately to make different value judgments about it.

To sum up this section, then, it is little wonder that over the decades many evaluation reports have been ignored or pushed aside, because they focused on an issue or a function that was not the main concern of the *stakeholders*—that is, they gathered information that was not relevant to the real decisions that the stakeholders were interested in making. *And so, it always behooves the evaluator to ask—at the beginning of the assignment—questions such as the following*:

> What individuals or groups are stakeholders in the program, that is, who stands to gain or lose from the success or failure of the program?
>
> What will these stakeholders want to know about the program?
>
> Is a decision concerning the program about to be made?
>
> What are the options being considered—and when is the decision likely to be made (and by whom will it be made)?

The likelihood that the information gathered in an evaluation will be relevant to the concerns that exist and to the decisions that are likely to be made, often can be increased if the evaluator establishes an "Advisory Board" containing (among others) representatives of all the stakeholder groups. Although useful, this approach does not always work; there have been cases where the fate of a program has been so politicized and the rhetoric surrounding its evaluation has become so heated, that advisory board members—and evaluators themselves—have been so intimidated that they have been fearful to depart from some "party line". A case in point—which I will not document in order to protect the innocent—concerns the evaluation of bilingual education programs in the USA in the 1980s. The article by Maria de Ibarrola again springs to mind here, for she describes a situation that was probably much too discordant for an advisory group to be able to function. On the other hand, in diverse, multicultural settings of the kind discussed by Schmelkes, often such groups can be very effective.

The Formative/Summative Distinction

It is time to return to Cronbach's essay of 1963. In it he made the point that, with respect to educational programs or educational interventions, there were three general aspects about which decisions often needed to be made—and therefore about which the evaluator could gather relevant information:

1). Course improvement: deciding what instructional materials and methods are satisfactory and where change is needed;
2). Decisions about individuals: identifying the needs of the pupil for the sake of planning his instruction, judging pupil merit for purposes of selection and grouping, acquainting the pupil with his own progress and deficiencies; and
3). Administrative regulation: judging how good the school system is, how good individual teachers are, etc.

Cronbach's interest in this 1963 essay was in the first category, which is why the piece was titled "Course Improvement Through Evaluation". And given this emphasis on using evaluations to aid the making of decisions about how to improve the effectiveness of courses in the curriculum (and other educational programs and interventions), it is clear that Cronbach was an early advocate of what are now known as *formative evaluations*. It is worth making the point that there is delicious irony here, for Michael Scriven is widely (and justly) given credit for being the first to make very explicit (and name) the distinction between *formative* and *summative* evaluations. The distinction is often presented in the literature in this colorful and highly apt way: When the cook tastes the soup, this is a formative evaluation; when the customer tastes the soup, this is summative evaluation. (This wording is actually Bob Stake's and not Scriven's.)

More on Summative Evaluations

There comes a time in the life of a mature and well-established program when it seems to have grown stale, or when social changes or new intellectual developments suggest that the time has come to make serious updates or to replace the program entirely. Summative evaluation has an important role in this type of situation, for it can give an account of the "pluses" and "minuses" of the program—it "sums up" the overall costs and benefits (and deficiencies) and prepares the way for what is often called a "go/no go" decision. Overwhelmingly, summative evaluators have held the randomized controlled trial (RCT) as their methodological ideal, and Campbell and Stanley's classic *Experimental and Quasi-Experimental Designs for Research* has been their lay bible. This, in part, was the kind of thing that Lee Cronbach was referring to when he pointed out (in an earlier quotation) that evaluators had become "enthralled with technique."

Indeed, the summative model can be bewitching, and thus may be used when it is not appropriate; certainly during the late 1980s and 90s, the evaluation community still regarded carrying out large scale summative evaluations as more prestigious (more "sexy") than doing formative evaluations. An anecdote from the early days of the SEC illustrates some of the issues: Using some of its own scarce resources, plus international aid money, a developing country with a widely-spread rural population had built TV transmission towers so as to reach remote areas, and also had set-up TV viewing stations in settlements in these areas. Children and even adults were encouraged to come to these centers to watch specially produced literacy programs. Within several months of the TV programs being first broadcast, our group at Stanford was visited by members of an evaluation unit that had been established in this country -- a unit that seemed to be very powerful, as it reported directly to the country's president. The purpose of the visit was to get our input on the summative

evaluation they were planning. I still recall Cronbach's tone of voice as he asked whether, after several years of effort and the expenditure of vast amounts of money, it was likely that a decision would be taken to close the program down if it was not working well (and it would likely not be working well, because it would still be having "teething problems"). The answer, of course, was that such a decision was out of the question! So why, then, do you want to carry out a summative evaluation? The perplexed answer to this was that it was expected, and there was no alternative. The remainder of the visit was occupied with SEC members explaining that a formative evaluation—which would provide information that likely could lead to vast program improvements—was a much wiser investment of the nation's scarce resources (and would probably be much cheaper to conduct).

Nevertheless, it is still common to find evaluators who think of *all* evaluations in terms of the stereotypical model of summative evaluations done in an experimental mode. Cronbach and his SEC associates argued, years ago, that evaluators who thought this way were actually swearing blind allegiance to methodological commandments that have a religious tone to them:

- Thou shalt test the worth of a program whose goals are definite and whose plans have been fully worked out. Otherwise don't evaluate.
- Thou shalt compare. Compare the program that is of central interest with almost anything else, but compare!
- Thou shalt assign. Preferably, distribute subjects or institutions so as to make the comparison groups equivalent.
- Thou shalt measure goal attainment.
- Thy instruments shall be reliable.
- Thy procedures shall be objective.
- Thou shalt judge. Tell the client how good or bad the treatment is. (Cronbach and Associates, 1980, p. 215)

In addition, Cronbach's group gave some advice here:

> Much that is written on evaluation recommends some one "scientifically rigorous" plan. Evaluations should, however, take many forms, and less rigorous approaches have value in many circumstances.

Shavelson argues essentially the same points in his special issue article.

There is a final cautionary note about summative evaluations, especially ones that are focused upon large-scale programs, and that use an experimental design. Of necessity, such evaluations are conducted out in the field—in the real-life settings in which the relevant educational programs are running. And in real life, much can happen to derail the best laid plans of evaluators and researchers. Evaluators must be guided by a piece of folk wisdom: anything that can possibly go wrong is likely to go wrong! Partly for this reason, Cronbach and the members of the SEC gave this advice:

> It is better for an evaluative inquiry to launch a small fleet of studies than to put all its resources into a single approach. (Cronbach and Associates, 1980, p. 7)

A classic example of an evaluation that would have been much better off had this advice been followed (indeed, it is a case that should make any evaluator have second thoughts about undertaking a complex study incorporating an experimental design) is the evaluation of a trial (in multiple sites across the USA) of a program known as "educational performance contracting"

(Gramlich & Koshel, 1975). Teachers went on strike; a hurricane destroyed one of the field sites; pupils recorded obscenities on high-tech equipment being used in the program (and at another site the equipment was thrown out of upper-story windows of the school); delivery of the rewards to pupils—an integral part of the program at one site—was delayed by many months; and elsewhere many pupils failed to turn up for the post-testing session.
With hindsight it is clear that this program should have been given a formative rather than a summative evaluation—which brings us to the next topic.

More on Formative Evaluations

Over the years there has been a slowly growing realization that formative evaluations are sometimes (if not often) more useful and productive than summative evaluations—added to which there is more "bang for the buck", for their budgets usually are considerably smaller! One issue in dispute early on was whether formative evaluations were a *different type* of evaluation from summative. Scriven was quite adamant that the formative/summative distinction marked different *functions* of evaluation, not different *types*.

In contrast to Scriven, although Cronbach and the members of the SEC agreed that sometimes a summative evaluation could be used formatively, they also held that often formative evaluations seemed to be of a different type from summative. Not only were they frequently smaller in scale (and therefore cheaper), they often could be carried out much more quickly (in situations where timeliness was of the essence). Finally, they could often be less formal in design—it was not necessary for most of them to be based on the model of the randomized controlled trial! Rigor is not the sole province of randomized controlled experiments.

Formative evaluations are especially appropriate when a program is under development, or when it is in the early stages of implementation—for in such cases, decisions are being made about how to improve the program, how to deal with "glitches" or unexpected problems or difficulties or shortcomings. And in such cases, decisions need to be made promptly. A "small fleet" of "less rigorous" studies has much to recommend it here. (Without the guidance coming from formative evaluators, too often the program developer has had to rely on guesswork about how to improve the program.)

A classic illustrative example is the formative evaluation of the children's TV program *Sesame Street/Plaza Sesamo* (Cook et al., 1975). The original aim of this program was to use the medium of TV to provide underprivileged children with knowledge and skills they often lacked when entering schools—a deficit that caused them to fall further and further behind their more privileged classmates. The program developers decided to adopt a "magazine" format for each of the one-hour programs; that is, each program consisted of a number of brief scenes or skits, any one of which could be replaced if (in light of formative evaluation) it turned out to be ineffective. Each of these short skits involved several characters—a mix of live actors and puppets such as the famous "Miss Piggy" and the "Cookie Monster" (the aim of this was to fully engage children's attention). Furthermore, each one-hour program was "sponsored" by a couple of numbers and letters of the alphabet; so, for example, a program would be introduced with the announcement "Today's program is sponsored by the letters A and W; and by the number 3"; and these sponsors would appear several times during that particular program. So, then, the developers needed to make a number of crucial decisions along the following lines:

- What was the maximum number of sponsors that was conducive to learning?
- How many times should each sponsor appear during the hour, to produce the maximum learning?

- How many refresher appearances of these letters and numbers should there be in later programs?
- What was the maximum number of humans and puppets that should be on screen at the one time before these characters were distracting and learning decreased?
- What factors, such as type of voice, distracted viewers and decreased attention and learning?
- What was the optimum length of each skit?

Short, rather informal studies were run, with small numbers of children, to quickly determine the answers to these questions (the timeframe was hours or days, not weeks or months). In the course of these formative inquiries, it was found (unexpectedly) that the presence of an adult who merely viewed the program along with the children, increased attention and fostered greater learning—and this information prompted the developers to incorporate features into the skits that made the program of interest to adults (for example, jokes and satirical references, and the puppet characters themselves—Miss Piggy was as popular with adults as with children!).

Evaluation versus Scientific Research

There is this one last general issue to pursue—and I must begin with a confession about it. (It is a long one, stretching over the next two paragraphs!) Until recently, I was a strong supporter of the following view: Although the modern enterprise known as the evaluation of educational programs was a child of the field of educational research, there are some very important differences between mother and offspring. There is, of course, a family resemblance that remains—namely, that projects or inquiries undertaken in both the evaluation and research domains are expected to reach conclusions that are *well warranted*, that are supported by relevant evidence and argument, and where the process of inquiry has taken account of threats to validity. But the evaluation child is not the same as the research parent. Certainly it is difficult to pin down the precise differences, and there can be considerable overlap, but nevertheless they are different although related endeavors. There seems to be merit in the insight of Cronbach and Suppes that research is conclusion oriented while evaluation is decision oriented.

At the operational level there frequently seem to be differences between the two. Many (but not all) scientific research projects have been drawn-out affairs—think of the decades over which Darwin's research was spread, or Einstein's work on relativity—and crucially, in a research project, if more time is needed to pin down a conclusion, more time is taken. Many projects in applied science, and certainly in evaluation, have a short time-line that absolutely cannot be stretched (maybe the report is needed next month for the meeting of a legislative committee, or perhaps the start of the school year is looming). The project to develop the first atomic bomb (an exercise in applied science) was under severe time pressure, because every day of delay cost lives in battles in the Pacific. (Galileo, Newton, Boyle, Darwin, and others were under no comparable pressure!) Applied projects also usually have tight budgets, so that a project's design has to take cost into account; this is not to deny that scientific research is underfunded around the world—but if a research proposal is funded, its central elements are adequately financed (otherwise there is no point in doing it).

As the perspicuous reader will have surmised from my having made this confession, I no longer hold this view—or at least, not so firmly! It was not Michael Scriven's dismissing of the distinction that changed my mind:

> … attempts to distinguish research from evaluation—some identify six or eight dimensions—distort or stereotype one or the other. For example, it is often said that

> evaluations are aimed at conclusions that are 'particularistic' or ideographic rather than general or nomothetic, the latter supposedly being the goal of the scientific researcher. This is wrong in both directions…. (Scriven, 1991a, p. 159)

Rather, the decisive factor was the work of many of my colleagues reported in this volume. Consider for example the contributions by van der Berg, and Ercikan and her colleagues; these report evaluations, but they also are interesting and potentially important pieces of research. The same can be said of several other chapters. Good work is good work, and it has the potential to be useful in many contexts—a thought that is an appropriate launching point into the interesting chapters that follow.

References

Cook, T., Appleton, H., Conner, R. F., Shaffer, A., Tamkin, G., & Weber, S. J. (1975). *Sesame Street Revisited.* NY: Russell Sage Foundation.

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

Cronbach, L. J., and Associates. (1980). *Toward Reform of Program Evaluation.* San Francisco: Jossey Bass.

Gramlich, E., & Koshel, P. (1975). *Educational Performance Contracting.* Washington, D.C.: Brookings.

McLaughlin, M., & Phillips, D.C. (1991)). Eds. *Evaluation and Education: At Quarter Century.* Chicago: University of *Chicago Press/NSSE.*

Scriven, M. (1991a). *Evaluation Thesaurus.* Thousand Oaks CA: Sage.

Scriven, M. (1991b). Beyond formative and summative evaluation. In M. McLaughlin & D. C. Phillips (Eds.), *Evaluation and Education: At Quarter Century* (pp. 19-64). Chicago: University of Chicago Press/NSSE.

About the Author

D. C. Phillips
Stanford University
d.c.phillips@gmail.com
D. C. Phillips was born, educated, and began his professional life in Australia; he holds a B.Sc., B.Ed., M. Ed., and Ph.D. from the University of Melbourne. After teaching in high schools and at Monash University, he moved to Stanford University in the USA in 1974, where for a period he served as Associate Dean and later as Interim Dean of the School of Education, and where he is currently Professor Emeritus of Education and Philosophy. He is a philosopher of education and of social science, and has taught courses and also has published widely on the philosophers of science Popper, Kuhn and Lakatos; on philosophical issues in educational research and in program evaluation; on John Dewey and William James; and on social and psychological constructivism. For several years at Stanford he directed the Evaluation Training Program, and he also chaired a national Task Force representing eleven prominent Schools of Education that had received Spencer Foundation grants to make innovations to their doctoral-level research training programs. He is a Fellow of the IAE, and a member of the U.S. National Academy of Education, and has been a Fellow at the Center for Advanced Study in the Behavioral Sciences. Among his most recent publications are the *Encyclopedia of Educational Theory and Philosophy* (Sage; editor) and *A Companion to John Dewey's "Democracy and Education"* (University of Chicago Press).

About the Guest Editors

Lorin W. Anderson
University of South Carolina (Emeritus)
anderson.lorinw@gmail.com
Lorin W. Anderson is a Carolina Distinguished Professor Emeritus at the University of South Carolina, where he served on the faculty from August, 1973, until his retirement in August, 2006. During his tenure at the University he taught graduate courses in research design, classroom assessment, curriculum studies, and teacher effectiveness. He received his Ph.D. in Measurement, Evaluation, and Statistical Analysis from the University of Chicago, where he was a student of Benjamin S. Bloom. He holds a master's degree from the University of Minnesota and a bachelor's degree from Macalester College. Professor Anderson has authored and/or edited 18 books and has had 40 journal articles published. His most recognized and impactful works are *Increasing Teacher Effectiveness, Second Edition*, published by UNESCO in 2004, and *A Taxonomy of Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*, published by Pearson in 2001. He is a co-founder of the Center of Excellence for Preparing Teachers of Children of Poverty, which is celebrating its 14th anniversary this year. In addition, he has established a scholarship program for first-generation college students who plan to become teachers.

Maria de Ibarrola
Department of Educational Research, Center for Research and Advanced Studies
mdeibarrola@gmail.com
Maria de Ibarrola is a Professor and high-ranking National Researcher in Mexico, where since 1977 she has been a faculty-member in the Department of Educational Research at the Center for Research and Advanced Studies. Her undergraduate training was in sociology at the National Autonomous University of Mexico, and she also holds a master's degree in sociology from the

University of Montreal (Canada) and a doctorate from the Center for Research and Advanced Studies in Mexico. At the Center she leads a research program in the politics, institutions and actors that shape the relations between education and work; and with the agreement of her Center and the National Union of Educational Workers, for the years 1989-1998 she served as General Director of the Union's Foundation for the improvement of teachers' culture and training. Maria has served as President of the Mexican Council of Educational Research, and as an adviser to UNESCO and various regional and national bodies. She has published more than 50 research papers, 35 book chapters, and 20 books; and she is a Past-President of the International Academy of Education.

D. C. Phillips
Stanford University
d.c.phillips@gmail.com
D. C. Phillips was born, educated, and began his professional life in Australia; he holds a B.Sc., B.Ed., M. Ed., and Ph.D. from the University of Melbourne. After teaching in high schools and at Monash University, he moved to Stanford University in the USA in 1974, where for a period he served as Associate Dean and later as Interim Dean of the School of Education, and where he is currently Professor Emeritus of Education and Philosophy. He is a philosopher of education and of social science, and has taught courses and also has published widely on the philosophers of science Popper, Kuhn and Lakatos; on philosophical issues in educational research and in program evaluation; on John Dewey and William James; and on social and psychological constructivism. For several years at Stanford he directed the Evaluation Training Program, and he also chaired a national Task Force representing eleven prominent Schools of Education that had received Spencer Foundation grants to make innovations to their doctoral-level research training programs. He is a Fellow of the IAE, and a member of the U.S. National Academy of Education, and has been a Fellow at the Center for Advanced Study in the Behavioral Sciences. Among his most recent publications are the *Encyclopedia of Educational Theory and Philosophy* (Sage; editor) and *A Companion to John Dewey's "Democracy and Education"* (University of Chicago Press).

SPECIAL ISSUE
Historical and Contemporary Perspectives on Educational Evaluation

education policy analysis archives

Volume 26 Number 46 April 16, 2018 ISSN 1068-2341

Please send errata notes to Audrey Amrein-Beardsley at Audrey.beardsley@asu.edu

Join EPAA's Facebook community at https://www.facebook.com/EPAAAAPE and **Twitter feed** @epaa_aape.

arquivos analíticos de políticas educativas
conselho editorial