Evaluation has always been a most difficult task, and Scriven (1972a) has gone so far as to characterize it as requiring all human skills known. Such an awesome requirement suggests that, at least as a start, help be sought from the professional evaluation literature; however, Borich (1974) among others has noted that much of the literature has not been immediately applicable or has represented untested methods. Although several evaluation models have been created, they are at most value (not empirically) based, and there is little information about the relative efficacy of the different models (Smith & Murray, 1974). This apparent sparsity of specific and practical evaluation advice is understandable in light of the fact that contemporary educational evaluation is a relatively new discipline that many feel had its start with the publication of Scriven's 1967 paper entitled, "The Methodology of Evaluation," or, perhaps, Cronbach's 1963 article, "Course Improvement Through Evaluation." Since then, the educational evaluation literature has proliferated, with the result that an overwhelming amount of often-conflicting advice is now available, but in elusive and scattered documents. If high-quality evaluation in large-scale development settings is being sought, however, it is essential to be cognizant of what the professional evaluation literature has to offer and to be able to judiciously select appropriate resources.

Evaluation Resources + Practical Experience
Because of the unique setup of the Curriculum Research and Development Group (CRDG) of the University of Hawaii, it has been possible to evolve an evaluation approach that has attempted to use the best ideas from the professional literature combined with the more practical considerations that were based on actual experience. In some sense, evaluation at CRDG has been both easier and more difficult to carry out than most evaluations done elsewhere. The extra difficulty has been, in part, due to the complexity as well as to the large-scale (often statewide) implementation of the programs. On the other hand, because of concentrated resources and financial support, the evaluation activities have had a sort of support that does not often exist elsewhere, but it is only in combination with existing evaluation resources that such support can lead to quality evaluation.

Adaptation of models. Rather than stick to one model, the evaluation staff has found it useful to follow Worthen and Sander's (1973) advice, taking an eclectic approach and selecting useful aspects from each of the many models. One of the first comprehensive formative (to help improve the product or program) evaluation models published was that developed by Sanders and Cunningham (1973), who extended Scriven's (1967) work. An updated version of the Sanders-Cunningham model, which was presented in Borich (1974), included specific suggested techniques to be used at each stage of the formative evaluation. Whereas Sanders and Cunningham divided formative evaluation into four categories — predevelopmental, evaluation of objectives, interim, and product — Hess and Wright (1973) chose to characterize the stages as initiation, hot house, pilot test, field test, and public diffusion. These two models were combined with ideas from Stufflebeam (1968) and Scriven (1967, 1974a) to form the framework for the formative evaluation design (Villanueva-King, 1975) of the Hawaii English Project — Secondary which was still undergoing development.

The development of the elementary school version of the Hawaii English Program, on the other hand, has already been completed. Its summative (to see the effects of the program) evaluation design (Lai, 1975a), therefore, was based on a different combination of models.
Consistent with Stake (1970) and Scriven (1967), an attempt was made to evaluate (in terms of need) the program objectives themselves. The evaluation was designed to provide information to decisionmakers (Alkin, 1969; Stufflebeam, et al., 1971), interpretation and description (Stake, 1970), and judgment (Scriven, 1967). In terms of goal-free evaluation (GFE) advocated by Scriven (1972b), the goal-based evaluation (goals largely determined by curriculum developers) has been supplemented with program-goal-free evaluation (goals determined by other than developers), as well as a purer sort of GFE which is an attempt to be sensitive to side effects. Finally, Scriven's (1974b) summative checklist was used as a unifying construct.

The evaluation of the Foundational Approaches in Science Teaching (FAST) Program, another large-scale curriculum program, has combined the concept of Stake's (1967) model which looks at antecedents, transactions, and outcomes, with a decision research approach in which decision questions are the basis for evaluation. Integral to the FAST evaluation is the assessment of the curriculum theories on which the program is based. This intrinsic (Scriven, 1967) evaluation is complemented by an extensive array of formal, informal, observational, and goal-free instruments. A relatively unique focus of FAST, and hence its evaluation, is the curriculum support component, which includes teacher and administrator training, program implementation planning, field support services, and program monitoring and communication services. In addition, professional self-evaluation is stressed, whereby evaluation may continue long after any summative evaluation has been completed.

Each of the many other projects has taken a tailor-made approach to evaluation, attempting to combine the best of the theoretical and practical advice available. The necessity for uniqueness in evaluation approach becomes apparent from the following listing of some of the current projects: multicultural studies, mathematics, consumer education, bilingual education, and early childhood education.

Meta-evaluation (evaluation of the evaluation) approaches that have been adapted include Stufflebeam et. al.’s (1971) exhaustive list, Scriven’s (1974b) checklist, and Sanders and Nafziger’s (1975) guidelines.

Formative/Summative distinction. As might be expected, the Legislature and the Department of Education have been more interested in summative than in formative evaluation. Unfortunately, summative evaluation is often asked for immediately even though very little formative evaluation has taken place. This state of affairs is hardly conducive to carrying out summative evaluations. The formative/summative distinction in CRDG’s setting exhibits a pattern that is different from the distinction often made in the literature; for example, what would definitely be formative evaluation, say, in the Sanders-Cunningham (1973) schema, may be summative in that future funding will depend on whether the product is effective, regardless of the fact that the product has not yet been integrated into a complete program or curriculum. If only did formative evaluation (i.e., in order to revise the product), then financial support might become shaky. On the other hand, a purely summative approach would not take advantage of the potency of formative evaluation to help improve products. Attempts to accommodate such concerns have led to labels like “interim summative evaluation.” This type of evaluation is similar to that performed in the interim/external evaluation cell in the Sanders-Cunningham (1973) formative evaluation model; however, the term “summative” is a crucial difference. Borich (1974) and others have indicated that the formative/summative distinction is often not clear. At CRDG, not only is the demarcation line often blurred, but also certain evaluation stages are simultaneously both formative (to improve the product) and summative (to see if further financial support is justified).

In general, the curriculum developers have been somewhat uneasy about evaluation that is overly summative in nature, and it is important to emphasize that most of the evaluation efforts that concern them are geared toward a goal that developers as well as evaluators have: the improvement of the product; however, when the legislature and Department of Education ask
about evaluation, they almost always mean summative evaluation. Rather than simply report only summative results, it is recommended that formative results also be presented as evidence of attempts to improve the quality of the product. In the past, officials have been pleased to receive such information, but they had simply not thought of the attempts to improve the product as being related to “evaluation.”

**Sampling considerations.** Hawaii’s multicultural makeup poses challenging evaluation problems. For example, how does one draw a representative sample when almost a dozen ethnic groups have substantial representation in the population? Certain test questions may be biased against one group but not another. Mainland-developed tests can be (and often are) biased against islanders (Dykstra, et. al., 1975). With approximately 8 percent of our population being immigrants (the national figure is 2 percent), it is virtually impossible to make valid, controlled comparisons with how the rest of the United States is doing (e.g., on standardized achievement tests), although such comparisons are constantly being made by many educators, legislators, and evaluators. Similar problems undoubtedly exist elsewhere.

Because sampling is often done at the school level, it is essential to have relevant data that are relatively accessible on all schools. Socioeconomic (SES) data that are available include percentage of families on welfare, block statistics from the U.S. Census, occupations and salaries of heads of households, average cost of housing, transiency rates, etc. By having recent SES information, it is possible to quickly select randomly (stratified) a group of schools that either is representative of the State or satisfies a certain composition need. Sampling at the pupil level is also facilitated by an up-to-date file of individual demographic data.

**Selection and development of measures.** Perhaps in part because of Hawaii’s isolation, the need to be current with regard to the work being done elsewhere has been recognized. Measures are located, screened, selected, or rejected with the assistance of UCLA’s Center for the Study of Evaluation’s comprehensive critiques of existing tests, the Robinson and Shaver (1973) compendium of tests in the social psychological area, various test summaries put out by Educational Testing Service, the National Council of Teachers of English, Research for Better Schools, Buros, and others. Objectives and item banks that are used where helpful include those of IOX (UCLA), Westinghouse Learning Corporation, National Assessment, SOBER (Skager, 1974), and states like California and New York.

Norm-referenced, standardized achievement tests have been found to be generally inappropriate due to invalidity in the form of overgenerality, nonoverlapping of objectives, culture bias, low reliability for individual scores, questionable norms, confusing directions, and poor items (Dykstra, et. al., 1975).

A recent study (Lai, 1975b) on the Stanford Achievement Test used in Hawaii showed that the usual demographic (socioeconomic status) variables, like family income or parents’ education, accounted for as much as 86 percent of the variance in test scores, thus leaving very little that could be accounted for by a program that was being evaluated. The consumer who, on the other hand, wants to know how a certain group of children is doing in relation to the rest of the United States is asking for a basically impossible comparison. If, for example, one is comparing Hawaii fourth graders with mainland fourth graders, and using standardized achievement tests, the comparison is invalid unless at least the following hold true: (1) Hawaii’s socioeconomic status is similar to the mainland’s (there exists substantial evidence that this is not true); (2) the test measures only objectives that are considered worthwhile by a group that is an important part of the audience of the evaluation (this is important because for some subtests, one or two items constitute a year’s difference in grade equivalents); (3) test administration was of the same caliber as that used when the norming was carried out (there exists evidence that this is generally untrue); and (4) the norms are up-to-date as well as generally valid.
Tests that measure objectives relevant to an evaluation often have to be created by the evaluation staff. As expected, test development has been difficult, lengthy, costly, and sometimes not doable. Because of the costs and time involved, it has proved best to keep the number of tests being developed to a minimum. It is better to get reasonably good data in a few areas than a lot of questionable data from many areas. A lean data approach to evaluation is becoming increasingly popular in the literature (e.g., Popham, 1972).

In matters of test reliability or internal consistency, the evaluation staff has had total responsibility. However, in terms of validity, a content analysis, preferably based on some theory, performed by the program staff has been more likely to improve test validity than anything the evaluation staff does could do. Less reliance on statistics and more emphasis on content represents an approach advocated by Guttman (1976).

**Political Concerns**

Where community or legislative involvement in the selection of objectives is important, it is often useful and more effective to present interested parties with actual items and ask them to decide whether or not they agree with the item (rather than the more general objective) as being a valid measure of success, failure or quality. Such "sign-offs" before data are collected, however, do not guarantee the lack of future complaints by the various audiences (as experienced by Klein et al., 1975). Action at various levels of the school bureaucracy can sometimes be facilitated by liaison personnel who have connections in the necessary places. Such a person acts like an ombudsperson who can help oil the wheels for evaluations.

Evaluation reports must be written with specific audiences in mind. The usual thick technical document has not been a successful means of conveying evaluation results; furthermore, because of turnover in staff, combined with fallible memories, current evaluations must continually reiterate past evaluation results.

Finally, it seems that no matter how frequently and strongly statements are qualified (e.g., the reader is cautioned not to overgeneralize or make invalid comparisons), when the news media or public officials get hold of the results, abuses in dissemination occur. Two ways to alleviate this situation are (1) present the results along with all the necessary qualifications in an education session for the news media and public officials, and (2) include the specific cautions in the summary section of the report so that even those who read just the minimal amount will have been informed about important limitations of the evaluation.

**Future Evaluation**

Although there may be some disagreement with regard to what has happened to date in the field of evaluation, an evolutionary process theoretically results in a strengthening of the object that is evolving, and evaluation is evolving rapidly with respect to both time and breadth. The resulting discipline of contemporary evaluation is not easily mastered. Although the job of today's evaluators is far more complex than that of their predecessors, there is at the same time a real potential for doing better evaluation. That potential can be realized only if evaluation resources are skillfully combined with practical experience to produce a valid method of doing contemporary evaluation. If either familiarity with available evaluation resources or practical experience is lacking, then there is little hope for the attainment of quality evaluation in large-scale development settings.

**References**


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

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

Morris Lai has been a Senior Evaluator at the Curriculum Research and Development Group, University of Hawaii, for the past three years. He has designed evaluations of numerous programs, including the Hawaii English Program, the Bilingual/Bicultural Program, Title I of ESEA, and the Lau Desegregation Center. Prior to coming to Hawaii, he spent four years in evaluation and research at the Far West Regional Laboratory in San Francisco.